The Bracket Geometry of Statistics

Uncovering the Foundations of Hamiltonian Monte Carlo, Measure-Constraints Preserving Flows and Diffusions, Geometric MCMC, and Stein Operators

A thesis presented for the degree of
Doctor of Philosophy of Imperial College London
by,

Alessandro Andrea Barp

Department of Mathematics
Imperial College
180 Queen's Gate, London SW7 2BZ

May 2020

Abstract

In this thesis we build a geometric theory of Hamiltonian Monte Carlo, with an emphasis on symmetries and its bracket generalisations, construct the canonical geometry of smooth measures and Stein operators, and derive the complete recipe of measure-constraints preserving dynamics and diffusions on arbitrary manifolds.

Specifically, we will explain the central role played by mechanics with symmetries to obtain efficient numerical integrators, and provide a general method to construct explicit integrators for HMC on geodesic orbit manifolds via symplectic reduction. Following ideas developed by Maxwell, Volterra, Poincaré, de Rham, Koszul, Dufour, Weinstein, and others, we will then show that any smooth distribution generates considerable geometric content, including "musical" isomorphisms between multi-vector fields and twisted differential forms, and a boundary operator - the rotationnel, which, in particular, engenders the canonical Stein operator. We then introduce the "bracket formalism" and its induced mechanics, a generalisation of Poisson mechanics and gradient flows that provides a general mechanism to associate unnormalised probability densities to flows depending on the score pointwise. Most importantly, we will characterise all measure-constraints preserving flows on arbitrary manifolds, showing the intimate relation between measure-preserving Nambu mechanics and closed twisted forms. Our results are canonical. As a special case we obtain the characterisation of measure-preserving bracket mechanical systems and measure-preserving diffusions, thus explaining and extending to manifolds the complete recipe of SGMCMC.
We will discuss the geometry of Stein operators and extend the density approach by showing these are simply a reformulation of the exterior derivative on twisted forms satisfying Stokes' theorem. Combining the canonical Stein operator with brackets allows us to naturally recover the Riemannian and diffusion Stein operators as special cases. Finally, we shall introduce the minimum Stein discrepancy estimators, which provide a unifying perspective of parameter inference based on score matching, contrastive divergence, and minimum probability flow.

Acknowledgements

I would like to warmly thank my supervisor Mark Girolami, for his audacity to recruit someone that did not know the definition of a random variable on a statistics PhD, as well as for the freedom and opportunities he granted me. I am also grateful to Tony Kennedy for his patience, advices, and continuous support that greatly facilitated my journey in the land of HMC.

I would also like to express gratitude to Matthew Turner, Richard G. Morris, Daniel Ueltschi, and Gareth Alexander for introducing me to mathematical physics research, as well as Erlend S. Riis, and Paul Druce who played an important role in the construction of my mathematical knowledge, and in particular Dominic Bunnett - our philosophical debates enlightened me.

I also express my gratitude to the people that have contributed to this PhD, in particular Imperial College, the Alan Turing Institute, Chris Oates, So Takao, Lester Mackey, Tony Lelièvre, Houman Owhadi, Darryl Holm, and Michael Betancourt, for stimulating discussions, as well as Andrew Duncan for his patience, and François-Xavier Briol for introducing me to the world of statistics and many enjoyable moments.

Finally, I would like to thank my parents and family for their support and encouragement, the Little legends, and my school teachers, in particular Domenico Sgobba and Mazzeo, as well as my PhD examiners, Prof. Grigoris Pavliotis and Prof. Mathias Rousset, for an enjoyable viva examination and very useful comments that improved chapter 4.

Dedication

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline. The only exceptions are the numerical experiments in §6.8, and the plots in §3 which were done by François-Xavier Briol and Tony Kennedy respectively.

Alessandro Barp

Copyright

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law

Contents

Abstract **1**

Acknowledgements **3**

1 Introduction **18**

1.1 Motivations and Contributions 18

1.2 Publications 22

2 Sampling and Estimation of Probability Distributions **25**

2.1 Sampling and Numerical Integration 25

2.1.1 Mechanics 29

2.1.2 Hamiltonian Monte Carlo on Euclidean Space 30

2.1.3 Numerical Integration of Hamiltonian Flow 31

2.2 Estimation of Probability Measures 34

2.2.1 M, Z, and Wasserstein Distance Estimators 36

3 Hamiltonian Monte Carlo on Manifolds **40**

3.1 Measures and Geodesics on Manifolds 42

3.1.1 Haar Measures 43

3.1.2 Uniform Measures 44

3.1.3 Geodesic Flow 45
3.2 Measure-Preserving Flows with Mechanics 47
3.2.1 Dynamical Systems 47
3.2.2 Hamiltonian Mechanics 49
3.2.3 Hamiltonian Mechanics on $T\mathcal{M}$: Riemannian Newton Equations 52
3.2.4 Lagrangian Mechanics 54
3.3 Construction of Hamiltonian System Adapted to Target Distribution 55
3.3.1 Symplectic Structures on Fibre Bundles 56
3.3.2 Riemannian Newton System 58
3.3.3 Lie Group Hamiltonian System 58
3.4 The Hamiltonian Monte Carlo Algorithm 59
3.4.1 Sampling Canonical Distribution and Target Measure 59
3.4.2 Standard Implementations of HMC 61
3.5 Mechanical Integrators and Shadow Hamiltonians 65
3.5.1 Geodesic Integrators on $T\mathcal{M}$ 66
3.5.2 HMC on Manifolds with Holonomic Constraints 67
3.5.3 Generating Functions and Symplectic Integrators 71
3.5.4 Modified Equations and Shadow Hamiltonians 72
3.6 Hamiltonian Monte Carlo on Embedded Manifolds 75
3.6.1 Isometries are Symplectomorphisms 76
3.6.2 HMC via Embeddings 77
3.6.3 Matrix Groups 80
3.6.4 HMC on Submanifolds of $\mathbb{R}^\ell$ with Geodesic Integrator 82
3.6.5 Open Submanifolds of Vector Spaces 83

3.7 HMC via Riemannian Submersions . . . 84
3.7.1 $\tau : \mathbf{S}(n,k) \to \mathbf{G}(n,k)$ . . . 86
3.7.2 Hopf Fibration $\tau : S^{2n+1} \to \mathbb{CP}^n$ . . . 86
3.8 HMC on Shape Spaces by Reduced Dynamics . . . 86
3.8.1 Noether's Theorem and Reduced Dynamics . . . 88
3.8.2 Homogeneous Manifolds . . . 93
3.8.3 Naturally Reductive Homogeneous Spaces . . . 93
3.9 Hamiltonian Monte Carlo On Lie Groups . . . 99
3.9.1 Hamiltonian Monte Carlo on Lie Groups . . . 100
3.9.2 Geodesics . . . 105
3.9.3 Proofs . . . 107

4 The Bracket Geometry of Measure-Preserving Systems, and the Foundations of HMC 111

4.1 Diffusions on Euclidean Space . . . 113
4.1.1 Langevin Processes and Irreversible HMC . . . 116
4.1.2 MALA . . . 117
4.2 The Geometry and Dynamics Generated by Smooth Measures and the Bracket Formalism . . . 117
4.2.1 The Geometry and Dynamics of Measures . . . 118
4.2.2 The Bracket Formalism . . . 122
4.3 The Geometric Mechanics Foundations of HMC . . . 128
4.3.1 Analysis of the HMC Desiderata . . . 129
4.3.2 Shadow of Symplectic Integrators and HMC . . . 137
4.3.3 Time-Reversibility and Bracket MCMC . . . 139

4.4 Measure-Preserving Diffusions . . . 141
4.4.1 Bracket diffusion with forcing . . . 142
4.4.2 Completeness of the $\mathcal{A}$-Diffusions . . . 146
4.4.3 $\mathcal{A}$-Diffusions on Riemannian Manifolds . . . 147
4.4.4 Bracket Diffusion on Flat Euclidean Space . . . 149
4.4.5 Dissipative Form of $\mathcal{A}$-diffusions and Hamiltonian Noise . . . 150
4.5 Langevin Diffusions and Irreversible MCMC algorithms . . . 153
4.5.1 Langevin Pairs and Riemannian Newton Diffusions . . . 154
4.5.2 iLMCMC Algorithm . . . 156
4.6 The Potential Theory of Measures and Complete Recipe of Measure-Constraints Preserving Flows . . . 159
4.6.1 Potential Theory of Measures . . . 161
4.6.2 Nambu Mechanics and Measure-Constraints Preserving Flows . . . 163
4.6.3 History of Measure-Preserving Diffusions . . . 165
4.6.4 The Hamiltonian Conspiracy . . . 167

5 Statistical Models, Reproducing Kernel Hilbert Space, Gradient Descent, and the Geometry of Stein Operators 169

5.1 Statistical Divergences and Information Metric . . . 170
5.1.1 Statistical Models and the Fisher Riemannian Metric . . . 170
5.1.2 Parametrised Statistical Manifolds . . . 171
5.1.3 Divergence Functions . . . 172
5.1.4 HMC on Statistical Manifolds . . . 173
5.2 Natural Gradient Descent . . . 174
5.3 Reproducing Kernel Hilbert Spaces . . . 177

5.4 Stein Operators from Generators of Diffusions . . . 179
5.4.1 Stein's Method . . . 180
5.4.2 Diffusion Stein Operator . . . 181
5.5 Representers and Kernel Stein Discrepancy . . . 182

6 Minimum Stein Discrepancy Estimators for Unnormalised Models 186

6.1 Score Matching . . . 189
6.2 Minimum Stein Discrepancy Estimators . . . 191
6.3 Example 1: Diffusion Kernel Stein Discrepancy Estimators . . . 192
6.4 Example 2: Diffusion Score Matching Estimators . . . 193
6.5 Further Examples: Contrastive Divergence and Minimum Probability Flow . . . 195
6.6 Implementing Minimum SD Estimators: Stochastic Riemannian Gradient Descent 196
6.7 Theoretical Properties for Minimum Stein Discrepancy Estimators . . . 197
6.8 Numerical Experiments . . . 200
6.8.1 Rough densities: the symmetric Bessel distributions . . . 200
6.8.2 Heavy-tailed distributions: the non-standardised student-t distributions . 201
6.8.3 Robust estimators for light-tailed distributions: the generalised Gamma distributions . . . 202
6.8.4 Efficient estimators for a simple unnormalised model . . . 203
6.8.5 Efficiency of Minimum SD Estimators for Scale Parameters of Symmetric Bessel distributions . . . 203
6.8.6 Bias Robustness of Minimum SD Estimators for the Symmetric Bessel and Non-standardised Student-t Distributions . . . 204
6.9 Geometry of Stein Operators . . . 204
6.10 Bracket-Induced Stein Discrepancies . . . 211
6.11 Derivations on Diffusion Stein Discrepancies . . . 213

6.11.1 Diffusion Kernel Stein Discrepancies . . . 214
6.11.2 Diffusion Kernel Stein Discrepancies as Statistical Divergences . . . 216
6.11.3 Diffusion Score Matching Derivations . . . 217
6.12 Information Semi-Metrics of Minimum Stein Discrepancy Estimators . . . 221
6.12.1 Proof of 39: Information Semi-Metric of Diffusion Kernel Stein Discrepancy 221
6.12.2 Proof of 40: Information Semi-Metric of Diffusion Score Matching . . . 222
6.13 Proofs of Consistency and Asymptotic Normality for minimum Stein Discrepancy Estimators . . . 223
6.13.1 Diffusion Kernel Stein Discrepancies . . . 223
6.13.2 Diffusion Score Matching . . . 227
6.13.3 Strong Consistency and Central Limit Theorems for Exponential Families 229
6.14 Proofs of Robustness of Minimum Stein Discrepancy Estimators . . . 235
6.14.1 Robustness of Diffusion Kernel Stein Discrepancy . . . 235
6.14.2 Robustness of Diffusion Score Matching . . . 236

7 Conclusion **238**

Bibliography **239**

A Appendix **270**

A.1 Vector Calculus . . . 270
A.2 Convergence of Measures . . . 270
A.3 Intuitive Description of Geometric Objects . . . 271
A.3.1 Manifolds, Vector Fields, Flows and Differential Forms . . . 272
A.3.2 Riemannian Metrics and Connections . . . 274
A.4 Further Background Material . . . 279

A.4.1 Origins of HMC . . . 279
A.4.2 Shadow HMC . . . 282
A.4.3 Approximate-Gradient HMC for Statistical Applications . . . 284
A.4.4 Hilbert Space HMC . . . 286
A.4.5 Ergodicity of HMC . . . 287
A.4.6 Generative Models, and Likelihood-free Inference . . . 288
A.4.7 Learning the Score . . . 290
A.4.8 Sinkhorn Statistical divergences . . . 291
A.4.9 Exponential Family . . . 292
A.4.10 Discrete Space Diffusion and Markov Chains . . . 293
A.5 Further Material HMC . . . 294
A.5.1 Riemannian Measure . . . 294
A.5.2 Retractions . . . 295
A.5.3 Principle of Stationary Action . . . 295
A.5.4 Motion in a Potential Field Vs Free Motion . . . 296
A.5.5 Discrete Variational Principle . . . 297
A.5.6 Hamilton-Jacobi PDE . . . 299
A.5.7 Additional Proofs HMC . . . 303
A.6 Additional Material Diffusions . . . 309
A.6.1 Overdamped Langevin Processes . . . 309
A.7 Additional Material on DKSD . . . 310
A.7.1 $\mathcal{B}$-KSD . . . 310
A.7.2 The Stein Kernel . . . 311
A.7.3 Special Cases of Diffusion Kernel Stein Discrepancy . . . 312

A.8 Minimum MMD Estimators for Generative Models 314
A.8.1 The Maximum Mean Discrepancy Statistical Manifold 315
A.8.2 The Natural Gradient Flow induced by MMD 316
A.8.3 Theory of Minimum MMD estimators 317

List of Tables

1.1 Notation used throughout the thesis. 24

3.1 Properties of integrators 60

3.2 Some examples of naturally reductive manifolds, and their decomposition. . . . 95

4.1 A summary of some standard mechanical systems and their properties. 127

4.2 A summary of some of the results 165

A.1 Hamiltonian system on $\mathcal{G} \times \mathfrak{g}$ and its image under the representation ψ. The equations of motion on $\mathcal{G}\times\mathfrak{g}$ are given by (A.12), and their image on $\mathrm{GL}(n)\times\mathbb{R}^{n\times n}$ satisfies (A.13). 306

List of Figures

3.1 *Energy level sets* 61

3.2 *A typical geodesic integrator alternates between geodesic steps, and vertical gradient steps that maintain the position but changes the direction. The figure illustrates the impact of the choice of numerical trajectory length.* 66

3.3 Example of reduced trajectory for a mechanical system with SO(2) symmetry on $T^*\,\mathrm{SO}(3)$; by symmetry, if initially $J = 0$ then the system remains on the level set $J = 0$ and the reduced trajectory is a Hamiltonian dynamics on the sphere $\mathrm{SO}(3)/\,\mathrm{SO}(2)$. 91

3.4 20 HMC trajectories, chronologically coloured blobs, the trajectories are coloured proportionally to the values of δH, with potential $V(x, y, z) = yz^2 \exp(x^2)$, $\delta t = 0.1$, $\tau = 2$. 97

3.5 *HMC via reduction by symmetry for a well-known distribution* 98

6.1 *Minimum SD Estimators for the Symmetric Bessel Distribution.* We consider the case where $\theta_1^* = 0$ and $\theta_2^* = 1$ and $n = 500$ for a range of smoothness parameter values s in $d = 1$. 201

6.2 *Minimum SD Estimators for Non-standardised Student-t Distributions.* We consider a student-t problem with $\nu = 5, \theta_1^* = 25, \theta_2^* = 10$ and $n = 300$. 201

6.3 *Minimum SD Estimators for Generalised Gamma Distributions under Corruption.* We consider the case where $\theta_1^* = 0$ and $\theta_2^* = 2$ (left and middle) or $\theta_2^* = 5$ (right). Here $n = 300$. 202

6.4 *Estimators for a Simple Intractable Model* 203

6.5 *Minimum SD Estimators for the Scale of a Symmetric Bessel Distribution.* We consider the case where $\theta_1^* = 0$ and $\theta_2^* = 1$ and $n = 500$ for a range of smoothness parameter values s in $d = 1$. 203

6.6 *The Robustness of Minimum SD Estimators for the Symmetric Bessel and Student-t Distributions. Left:* Student-t distribution. *Right:* Symmetric Bessel distribution. 204

A.1 Arc length on the sphere between the points (θ, ϕ) and $(\theta + d\theta, \phi + d\phi)$. 275

A.2 LHS: constant vector field wrt $\mathrm{d}r^2 + r^2\mathrm{d}\theta^2$. RHS: constant vector field wrt $\mathrm{d}r^2 + \mathrm{d}\theta^2$ 276

A.3 The geodesics on S^2 are the great circles. The tangent vector to a great circle, when viewed as a vector in $\mathbb{R}^3$, varies only along the normal to S^2. 278

Chapter 1

Introduction

1.1 Motivations and Contributions

This thesis contributes to the geometrisation of statistics, and, in particular, of measure-preserving systems which play a central role in many areas of statistics, machine learning, and physics. Specifically, the main objectives of this thesis are (i) construct a unified geometric theory of Hamiltonian/Hybrid Monte Carlo (HMC), discuss the role of symmetries in its implementation, and obtain a clear understanding of the statistical significance of properties of Hamiltonian mechanics; (ii) develop the theory of smooth measures and brackets for statistical purposes; (iii) characterise measure-preserving systems on manifolds (dynamics, bracket mechanics, and diffusions); (iv) extend the density approach of Stein operators and diffusion Stein operators to manifolds, and explain their relation; (v) construct minimum Stein discrepancy estimators that unify many previous methods for parameter inference, and establish the consistency, asymptotic normality, and robustness properties of two important subclasses of estimators based on diffusion kernel Stein discrepancy and diffusion score matching.

I was first introduced to HMC during the first year of my PhD, at a time when we were still unsure what exactly would be the focus of my research. As I had a background in mathematical physics, mechanics-based sampling schemes seemed like an interesting idea. Going over the HMC statistical literature, I was surprised by the absence of discussion regarding Hamiltonian mechanics and its geometry - it seemed the vast geometric mechanics literature was completely ignored. Indeed - even the theory of geometric integration is only superficially used (Lagrangian mechanics and variational integrators are essentially never mentioned in statistics), despite the fact that the main step of HMC consists in numerically integrating a mechanical trajectory. I later realised that there were in fact three HMC literatures, one from lattice

quantum chromodynamics (QCD), one from molecular dynamics, and the statistical one; with very little collaboration between these areas of research in spite of a common aim. Listening to conversations between famous researchers such as Tony Kennedy, Mark Girolami, Michael Betancourt, and Ben Leimkuhler, it quickly became evident that the reason for this was that most researchers in one area of application simply did not know what was happening in the other areas. Research in geometric mechanics, geometric integration, lattice QCD, statistics and molecular dynamics have thus progressed quite separately. This is probably due to the different jargon/notation employed, the vastness of the literature, but also results from more fundamental disagreements about elementary notions such as "phase space" and "mechanics". Indeed, as I will discuss, the lattice QCD literature has made up its own "mechanics" by constructing explicitly a symplectic structure from the Maurer–Cartan frame, and this fact caused a lot of confusion between Tony Kennedy and I when we were trying to build an analogous mechanics on symmetric spaces (and also between Tony Kennedy and Michael Betancourt concerning the role of fibre disintegration). As I will show, this symplectic structure simply corresponds to the left-trivialisation of canonical mechanics, though it is interesting to note that the Maurer–Cartan expansion of the symplectic structure seems to really originate in the HMC literature rather than the geometric mechanics one - the first analogous expansion I have been able to find in the latter is due to Dmitri Alekseevsky et al. [11].

In chapter §3 I will aim to provide a comprehensive theory of HMC on arbitrary manifolds, and discuss its implementation on the various classes of manifolds that arise in applications, in order to correct some false claims in the statistical literature which have caused some confusion. All the theory developed therein will be coordinate-invariant to show that, by combining the theory of fibre integration with geometric mechanics, one never needs to use coordinates when analysing HMC, and the geometric construction lifts several misconceptions. I will discuss in detail the role of symmetries, which allows us to construct explicit numerical integrators on any geodesic orbit manifold using the Lie group exponential. The resulting algorithm generalises [187, 186, 66, 162, 160, 202, 161]. Moreover I will uncover the relation between many variants of HMC, including the original HMC method used in lattice QCD, magnetic HMC, Lagrangian Monte Carlo, embedded HMC, and geodesic Monte Carlo: these "variations" of HMC simply correspond to the standard HMC scheme up to a symplectic transformation (symplectomorphism, symplectic embedding, or symplectic reduction).

In §4 I will carefully analyse the geometry of measure-preserving dynamics, mechanics, and diffusions. Indeed Radon measures on manifolds that are absolutely continuous with respect to the canonical null-sets of the atlas may be equivalently viewed as one densities (tensor densities of weight one). This geometric characterisation of measures allows us to construct a boundary operator, the rotationnel (or curl), that can be used to characterise measure-preserving dynamics. Specifically, any positive one density P defines an isomorphism between (twisted)

differential forms and multi-vector fields (a generalisation of the Hodge star isomorphism), which turns the exterior derivative on forms into a derivative curl_P associated to P. Any P-preserving dynamics may then be shown to be a locally curl vector field, that locally looks like $\mathrm{curl}_P(\mathcal{A})$ for some bi-vector field $\mathcal{A}$. I will also introduce the "bracket formalism" to develop a coherent theory of score based mechanical systems, which include gradient flows and Poisson mechanics, and provide a characterisation of the mechanical systems that preserve a smooth measure. Combining the geometric theory of measures with the bracket formalism allowed me to obtain a geometric mechanics theory of measure-preserving diffusions and HMC-like samplers, in which a generalisation of the modular vector field introduced by Weinstein for Poisson systems plays a particularly important role. For example I will show that any P-preserving diffusion has the form

$$\mathrm{d}Z_t = \mathrm{curl}_P(\mathcal{A})\mathrm{d}t + \tfrac{1}{2}\mathrm{div}_P(Y_i)Y_i\mathrm{d}t + Y_i \circ \mathrm{d}W_t^i,$$

for some vector fields (Y_i) and bi-vector field $\mathcal{A}$, up to an element of the first homology group of P. This important result improves and generalises to arbitrary manifolds the complete recipe of stochastic gradient MCMC which was derived in the popular paper [242]. Most importantly the derivation is constructive and is another demonstration of the power of geometry. The proof of [242] on Euclidean space relied on Fourier transforms, which are notoriously complex objects on manifolds which prevented its generalisation. Instead, the proof I provide shows that $\mathcal{A}$ may be explicitly constructed by Poincaré lemma via the (musical) isomorphisms defined by P - just as we usually construct the vector potential of the electromagnetic field. The result does not introduce any additional structure on the manifold, such as a Riemannian metric or connection. Furthermore, when $\mathcal{A}$ is viewed as a bracket, it generates its own mechanical system which maps log-densities (or Hamiltonians) to vector fields, $\mathcal{A} : H \mapsto X_H^{\mathcal{A}} = -i_{\mathrm{d}H}\mathcal{A}$, and if the target P takes the form $e^{-V}\mu_{\mathcal{M}}$ then the above diffusion naturally splits

$$\begin{aligned}\mathrm{d}Z_t = &\overbrace{\Big(\underbrace{\mathrm{curl}_{\mu_{\mathcal{M}}}(\mathcal{A})}_{\text{volume-preserving}} + \underbrace{X_V^{\mathcal{A}}}_{\text{density-preserving}}\Big)\mathrm{d}t}^{\text{local potential of Fokker–Planck current}} + \overbrace{P^\sharp(\gamma)\mathrm{d}t}^{\text{topological obstruction}} \\ &\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{conservative } L^2(P)\text{-antisymmetric drift}} \\ &+ \Big(\underbrace{-\tfrac{1}{2}Y_i(V)Y_i}_{\text{density-dissipative drift}} + \overbrace{\underbrace{\tfrac{1}{2}\mathrm{div}_{\mu_{\mathcal{M}}}(Y_i)Y_i}_{\text{volume-dissipative drift}}}^{\text{volume-preserving } L^2(\mu_{\mathcal{M}})\text{-symmetric diffusion}}\Big)\mathrm{d}t + \underbrace{Y_i \circ \mathrm{d}W_t^i}_{\text{Stratonovich noise}}. \\ &\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{L^2(P)\text{-symmetric fluctuation-dissipation balance}}\end{aligned} \tag{1.1}$$

Moreover, I will characterise the sub-class of P-preserving diffusions for which the volume contributions vanish: when the noise fields are (pointwise) linearly independent these have the

form

$$\mathrm{d}Z_t = \mathrm{curl}_{\mu_{\mathcal{M}}}(i_{\mathrm{d}V}\mathcal{V})\,\mathrm{d}t + \tfrac{1}{2}\mathrm{div}_{\mu_{\mathcal{M}}}(i_{\mathrm{d}V}\mathcal{A}_i)\mathrm{curl}_{\mu_{\mathcal{M}}}(\mathcal{A}_i)\mathrm{d}t + \mathrm{curl}_{\mu_{\mathcal{M}}}(\mathcal{A}_i)\circ \mathrm{d}W_t^i,$$

for 3-vector fields $\mathcal{V}$, and (noise) 2-vector fields $(\mathcal{A}_i)$. I will analyse the geometric mechanics of measure-preserving diffusions, for example deriving a general result for the generator to be "reversible up to a transformation", generalising the analogous result for the generator of Langevin diffusions and its reversibility up to momentum flip. I shall discuss the geometry of Langevin diffusions, and analyse their induced "irreversible HMC schemes" [288]. For example, on a Lie group with left-invariant metric I will show these these diffusions may be written as

$$\underbrace{\dot{q} = v_t^j e_j(q_t)}_{\text{symmetry reconstruction}} \quad , \quad \mathrm{d}v_t = \underbrace{\mathrm{ad}^{\top}_{v_t} v_t \mathrm{d}t}_{\text{Euler–Poincaré}} \underbrace{-G^{jk}e_j(V)(q_t)\xi_k \mathrm{d}t}_{\text{Potential flow}} \underbrace{-\frac{\beta}{2}v_t\mathrm{d}t + \mathrm{d}W_t}_{\text{OU}},$$

and extend the algorithm we had proposed in [17] (which was selected as one of the top 5 papers of the conference). Moreover I will derive a compressible bracket MCMC scheme extending [114] to manifolds, and show that unimodular Poisson manifolds share many of the key properties of symplectic manifolds in the context of implementing mechanics-based samplers with splitting integrators. I will then explain how the canonical geometry of P can be viewed as a *potential theory* of measure-preserving flows, that enables us to construct P-preserving flows that also preserve given constraints and symmetries. Effectively, the P-musical isomorphism identifies twisted $(\dim \mathcal{M} - \ell)$-forms with ℓ-multivector fields, which can be interpreted as Nambu mechanics, and closed twisted forms with P-preserving Nambu mechanics. Based on this observation, using the foliated de Rham cohomology I will derive a canonical characterisation of flows that preserve an arbitrary smooth positive measure and an arbitrary set of constraints on an arbitrary manifold.

Pairs associating a smooth measure and a bracket seem to enable the generalisation of many statistical schemes to manifolds. For example, I will show that these allow to break one of the myths in the construction of Stein operators: using results from de Rham cohomology we can show the density approach to construct Stein operators is valid on arbitrary manifold (rather than just $\mathbb{R}$), and there is often no need to introduce the generator approach. Indeed I will show that the two main Stein operators constructed from generators of measure-preserving diffusions, the Riemannian Stein and diffusions operators, naturally arise by associating the canonical Stein operator, which is really the divergence, with particular choices of brackets.

Finally in §6 I will analyse minimum Stein discrepancy estimators, which are particularly useful in the context of approximating target distributions only known through samples with differentiable unnormalised statistical models. I will show these generalise and improve on several important statistical estimators, and derive results about their convergence and robustness.

During my Ph.D. I made several other contributions. For example in [80] I developed (with Lester Mackey) the theory of a powerful method to approximate distributions using discrete measures; in [29] I constructed (with Chris Oates) a method to numerically integrate smooth distributions on manifolds; and in [62] I derived (with Andrew Duncan) several theoretical guarantees for inference using the Maximum Mean Discrepancy in the context of likelihood-free models. However I decided to only briefly mention these contributions in §5 and §6 in order to focus on the bracket and canonical geometry of measures and measure-constraints preserving flows, as these have important implications in many areas of statistics, and most importantly prove that when working with smooth target measures, one simply cannot ignore geometry.

1.2 Publications

Barp, A., Briol, F. X., Duncan, A., Girolami, M., & Mackey, L. (2019). Minimum stein discrepancy estimators. In Advances in Neural Information Processing Systems (pp. 12964-12976).

Barp, A., Kennedy, A., & Girolami, M. (2019). Hamiltonian monte carlo on symmetric and homogeneous spaces via symplectic reduction. arXiv preprint arXiv:1903.02699.

Barp, A., Oates, C., Porcu, E., & Girolami, M. (2018). A Riemannian-Stein kernel method. arXiv preprint arXiv:1810.04946.

Barp, A., Briol, F. X., Kennedy, A. D., & Girolami, M. (2018). Geometry and dynamics for Markov chain Monte Carlo. Annual Review of Statistics and Its Application, 5, 451-471.

Barp, A. (2019, August). Hamiltonian monte carlo on lie groups and constrained mechanics on homogeneous manifolds. In International Conference on Geometric Science of Information (pp. 665-675). Springer, Cham.

Arnaudon, **A., Barp**, A., & Takao, S. (2019, August). Irreversible Langevin MCMC on Lie Groups. In International Conference on Geometric Science of Information (pp. 171-179). Springer, Cham.

W. Y. Chen, **A. Barp**, F.-X. Briol, J. Gorham, M. Girolami, L. Mackey, and C. Oates. (2019) Stein point Markov chain Monte Carlo. In Proceedings of the 36th International Conference on Machine Learning.

Briol, F. X., **Barp, A.**, Duncan, A. B., & Girolami, M. (2019). Statistical inference for generative models with maximum mean discrepancy. arXiv preprint arXiv:1906.05944.

In preparation (and also discussed in this thesis):

Barp, A., Takao, S., Betancourt, M., & Girolami, M. (2020). The Bracket Geometry of Measure-Preserving Mechanical Systems and HMC *In preparation.*

Barp, A., Takao, S., Arnaudon, A., Betancourt, M., & Girolami, M. (2020). The Bracket Geometry of the Complete Recipe of Measure-Preserving Diffusions. *In preparation.*

Barp, A., Betancourt, M., & Girolami, M. (2020). The Bracket Geometry of Stein Operators. *In preparation.*

Barp, A., Betancourt, M., Lelièvre, T., Kennedy, T., & Girolami, M. (2020). Monte Carlo Sampling and Optimisation on Manifolds via Hamiltonian Dynamics. *In preparation.*

Pre-PhD:

Barp, A., Barp, E. G., Briol, F. X., & Ueltschi, D. (2015). A numerical study of the 3D random interchange and random loop models. Journal of Physics A: Mathematical and Theoretical, 48(34), 345002.

Configuration space (n dimensional)	$\mathcal{M}$ if manifold, $\mathcal{G}$ if Lie group
Reference Measure	$\mu_{\mathcal{M}}$ on $\mathcal{M}$, $\mathrm{d}q$ for Lebesgue measure on $\mathbb{R}^n$
Riemannian Measure	vol or $\mathrm{vol}_{\mathcal{M}}$
Target Measure	P when smooth, or $\mathbb{Q}$ when non-smooth
Target Density (smooth target)	ρ or p_∞, i.e., $P(\mathrm{d}q) \propto \rho(q)\mu_{\mathcal{M}}(\mathrm{d}q)$
Score	$\mathrm{d}\log\rho$
Phase Space	$(\mathcal{F}, \omega)$ symplectic Fibre Bundle $\pi : \mathcal{F} \to \mathcal{M}$
Hamiltonian Vector Field of H	$X_H : \mathcal{F} \to T\mathcal{F}$ or $\hat{H}$
Flow of Hamiltonian field X_H	$\Phi^{X_H} \equiv \Phi^H : \mathcal{F} \times \mathbb{R} \to \mathcal{F}, \quad \Phi^H_t : \mathcal{F} \to \mathcal{F}$
Evaluation of Vector fields	$X : q \mapsto X_q \equiv X(q) \in T_q\mathcal{F}$
Potential & Kinetic energies	$V : \mathcal{M} \to \mathbb{R}, V \equiv -\log\rho, \quad T : \mathcal{F} \to \mathbb{R}$
Numerical Integrator	Ψ or $\Psi_{\delta t}$; time step δt, trajectory length τ
Shadow Hamiltonian	$\mathcal{H}_\Psi$ or $\mathcal{H}_{\Psi_{\delta t}}$
Energy Difference	$\delta H \equiv H \circ \Psi - H, \quad \Delta H \equiv \mathcal{H}_\Psi - H$
Riemannian Metric on $\mathcal{M}$	$\langle \cdot, \cdot \rangle$, $\| \cdot \|^2 \equiv \langle \cdot, \cdot \rangle$, locally $G_{ij} \equiv \left\langle \partial_{q^i}, \partial_{q^j} \right\rangle$
Dual pairing on vector space W	$\langle \cdot, \cdot \rangle_* : W^* \times W \to \mathbb{R}$
Inner product on vector space W	$\langle \cdot, \cdot \rangle_W : W \times W \to \mathbb{R}$
Submersion on Homogeneous Space	$\tau : \mathcal{G} \to \mathcal{G}/\mathcal{K}$
Derivatives $f : \mathcal{M} \to \mathbb{R}$	$\mathrm{d}f$ exterior derivative, ∇f Riemannian gradient
Tangent Map $f : \mathcal{M} \to \mathcal{N}$	$\partial f : T\mathcal{M} \to T\mathcal{N}$, which is the Jacobian locally
Sections of Bundle $\mathcal{F} \to \mathcal{M}$	$\Gamma(\mathcal{F})$ or $\Gamma(\mathcal{M}, \mathcal{F})$, $\mathfrak{X}(\mathcal{M}) \equiv \Gamma(T\mathcal{M})$
Differential Forms	$\Omega(\mathcal{M})$, $\Omega^\ell(\mathcal{M})$ is space of ℓ-forms
Pullback/pushforward $f : \mathcal{M} \to \mathcal{N}$	$f^* : \Omega(\mathcal{N}) \to \Omega(\mathcal{M})$, $f_* : \Gamma(T\mathcal{M}) \to \Gamma(T\mathcal{N})$
Pushforward of Measures $f : \mathcal{M} \to \mathcal{N}$	$f_\sharp$ (equivalent to $(f^{-1})^*$ on one-densities)
Lie algebra $\mathfrak{g}_L$ of left-invariant fields	basis (e_i), dual basis (θ^i) of $\mathfrak{g}_L^*$
Lie algebra $\mathfrak{g} = T_1\mathcal{G}$	basis (T_i), general element ξ or X, Y
Action field on $\mathcal{G}/\mathcal{K}$ of $\xi \in \mathfrak{g}$	$\xi_{\mathcal{G}/\mathcal{K}}(q) \equiv \frac{\mathrm{d}}{\mathrm{d}t} e^{t\xi} \cdot q \big\|_{t=0} \in T_q(\mathcal{G}/\mathcal{K})$
Symmetric/antisymmetric part	$\Pi_{\mathcal{S}}(M) = \frac{1}{2}(M + M^\top)$, $\Pi_{\mathcal{A}}(M) = \frac{1}{2}(M - M^\top)$
$C_b(\mathcal{M})/C_c(\mathcal{M})$ functions $\mathcal{M} \to \mathbb{R}$	continuous bounded/compactly supported
$Z \sim P$	P-distributed random variable
Integration $\int f \mathrm{d}P$	$\mathbb{E}_P[f]$ or $P[f]$ or $\int fP$ when P smooth
Fibre integral on fibre bundles $\pi : \mathcal{F} \to \mathcal{M}$	$\dim(\mathcal{F}) - \dim(\mathcal{M}) = \ell$, $\int_{fib} : \Omega^{k+\ell}(\mathcal{F}) \to \Omega^k(\mathcal{M})$

Table 1.1: Notation used throughout the thesis.

Chapter 2

Sampling and Estimation of Probability Distributions

In this chapter we will introduce and briefly review two important statistical challenges that will be the focus of the thesis. Additional details can be found in §A.4. The first is the problem of sampling, and will be the topic of §2.1, where we will concentrate on Hamiltonian Monte Carlo (HMC) in Euclidean space, in preparation for chapters §3 and §4 which will analyse in detail the geometry of HMC and its generalisation. The second is the problem of estimation, and in §2.2 we will introduce the basic concepts of the theory of parameter inference, as well as some standard measures of discrepancy between probability distributions, in order to lay the groundwork for the theory of minimum Stein discrepancy estimators we develop in §5 and §6. The notation we will follow is summarised in table 1.1.

2.1 Sampling and Numerical Integration

One of the main challenges in statistics is the construction of methods that sample from probability distributions. The aim is to generate a sequence of random variables $X_i : \Omega \to \mathcal{M}$ for which the (random) measure $\frac{1}{m}\sum_{i=1}^{m} \delta_{X_i}$ converges to a given "target" distribution P on $\mathcal{M}$ in an appropriate sense. For example if the X_i are independent and P-distributed, Glivenko–Cantelli theorem guarantees (under appropriate conditions) that $\frac{1}{m}\sum_{i=1}^{m} \delta_{X_i}$ converges narrowly

to P almost surely, i.e.,

$$\frac{1}{m}\sum_{i=1}^{m} f(X_i) \to \int f \mathrm{d}P, \quad \forall f \in C_b(\mathcal{M}) \quad a.s.,$$

where $C_b(\mathcal{M})$ is the space of continuous and bounded real-valued functions on $\mathcal{M}$. This results illustrates the importance of sampling methods: once we have constructed the sequence (X_i) we can estimate the integral (or expectation) of a function $f : \mathcal{M} \to \mathbb{R}$

$$\mathbb{E}[f] \equiv \int f(x)\mathrm{d}P(x) \approx \frac{1}{m}\sum_{i}^{m} f(X_i). \tag{2.1}$$

Moreover if $\mu_{\mathcal{M}}$ is an arbitrary measure on $\mathcal{M}$, given a probability measure $P(\mathrm{d}x) = \rho(x)\mu_{\mathcal{M}}(\mathrm{d}x)$ with positive density $\rho > 0$, we can reduce the integration with respect to $\mu_{\mathcal{M}}$ to an expectation with respect to P: $\int f\mu_{\mathcal{M}} = \int (f/\rho)\mathrm{d}P = \mathbb{E}[f/\rho]$. Hence, sampling enables us to approximate integrals, a vital prerequisite for the applicability of scientific theories, where the predictive equations often require evaluating an integral that describes the average value of an observable.

For example in decision theory one is interested in evaluating the "risk" R associated to a decision d, which is the expected value of a loss function $R \equiv \int L\big(x, d(x)\big)\mathrm{d}P(x)$. In molecular systems consisting of a large number of particles, $\mathcal{M}$ is the phase space whose elements $(q, p) \in \mathcal{M}$ represent microstates, and the observables are functions of the microstates, such as the specific heat or bulk pressure [70]. The macroscopic properties are characterised by the expectation (and variance) of the observables with respect to the canonical distribution $\frac{1}{Z}e^{-\beta H}\mathrm{d}q\mathrm{d}p$, where H is the energy, $\beta > 0$, and $Z = \int e^{-\beta H}\mathrm{d}q\mathrm{d}p$ is the canonical partition function which also requires integrating over phase space. In the context of computational statistical physics expectations also arise in the problem of thermodynamic integration and free energy calculations [329]. In Bayesian statistics the target distribution is the posterior $P_{post}(\cdot|\omega)$ on a space $\mathcal{M}$ of parameters generated by some observation ω and a prior distribution P_{prior} on $\mathcal{M}$. The likelihood is a function $L(\omega|\cdot) : \mathcal{M} \to \mathbb{R}$ over parameters given by the Radon-Nikodym derivative

$$\frac{\mathrm{d}P_{post}(\cdot|\omega)}{\mathrm{d}P_{prior}}(\theta) = \frac{1}{Z}L(\omega|\theta), \qquad Z \equiv \int_{\mathcal{M}} L(\omega|\theta)P_{prior}(\mathrm{d}\theta).$$

The normalisation constant Z, which depends on the observation ω, is a quantity of interest representing the model evidence and is usually intractable. What makes the setting "Bayesian" is the fact the distribution P_{prior} is not associated to the distribution of a random observable but represents the degree of belief in the possible parameters. If $P_{data}(\cdot|\theta)$ denotes the distribution of the data (or observations) given a parameter θ, the prior and posterior predictive distributions

are computed by integrating with respect to the prior and posterior respectively

$$P_{prior,pred}(\mathrm{d}y) = \int_{\mathcal{M}} P_{data}(\mathrm{d}y|\theta) P_{prior}(\mathrm{d}\theta),$$
$$P_{post,pred}(\mathrm{d}y|\omega) = \int_{\mathcal{M}} P_{data}(\mathrm{d}y|\theta) P_{post}(\mathrm{d}\theta|\omega).$$

Finally, by the Feynman-Kac formula many PDEs have a "path integral" formulation: their solution can be written as a (conditional) expectation of a stochastic process.

The approximation of the expectation of f using the estimator in (2.1) is called the **Monte Carlo solution**. If $f : \mathcal{M} \to \mathbb{R}^n$, $X_i \sim P$ are independent, and $\mathbb{E}[f(X_1)^2] < \infty$, the central limit theorem (CLT) implies the fluctuations of the Monte Carlo estimator around the expected value $\mu \equiv \mathbb{E}[f(X_1)]$ converge in distribution to a Gaussian

$$\sqrt{m}\left(\frac{1}{m}\sum_{i=1}^{m} f(X_i) - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[(f(X_1)-\mu)(f(X_1)-\mu)^\top\right]\right).$$

The CLT shows that the rate of convergence of the Monte Carlo estimator is of order $\frac{1}{\sqrt{m}}$, independently of the dimension of $\mathcal{M}$. However in practice it is usually hard to generate i.i.d. sample of P. Indeed we often do not know the density ρ or can evaluate it only up to a normalisation factor. Moreover even when we do know ρ, there is no simple way to use knowledge of the evaluations of ρ at some points of $\mathcal{M}$ to device algorithms that generate samples from P. As described in 29.1 [243] this is true even when $\mathcal{M}$ is a finite set. When $\mathcal{M}$ is a high-dimensional space, we further need to take into account the concentration of measure phenomena according to which most of the probability mass will be concentrated in small (unknown) regions of $\mathcal{M}$. Even if a region of high probability mass is known, P is often multi-modal with its modes typically separated by large regions of negligible probability, making it difficult to establish general methods that explore the various modes.

Dropping on the requirement that the random variables (X_i) should be i.i.d., a more practical method to approximate expectations consists of defining a time-homogeneous Markov chain (X_i) that is ergodic and preserves P. The **Markov chain Monte Carlo** (MCMC) estimator is then built by running the chain for sufficiently long in order for it to reach the stationary regime, and then use the realisations of the Markov chain to form the estimate (2.1). If $(x, \mathrm{d}y) \mapsto \mathbb{P}(x \to \mathrm{d}y)$ denotes the **transition kernel** of the Markov chain, so

$$P_{X_{n+1}}(\mathrm{d}y) = \int \mathbb{P}(x \to \mathrm{d}y) P_{X_n}(\mathrm{d}x) = \int \mathbb{P}^n(x \to \mathrm{d}y) P_{X_1}(\mathrm{d}x),$$

where P_{X_i} is the distribution of X_i and $\mathbb{P}^n(x \to \mathrm{d}y)$ is the n-step transition kernel. We say $\mathbb{P}$

preserves P if

$$\int \mathbb{P}(x \to \mathrm{d}y)P(\mathrm{d}x) = P(\mathrm{d}y).$$

A sufficient condition for P to be an invariant distribution of the chain is that the chain is P-**reversible**,

$$\mathbb{P}(x \to \mathrm{d}y)P(\mathrm{d}x) = \mathbb{P}(y \to \mathrm{d}x)P(\mathrm{d}y).$$

One of the most widely-used method to generate chains preserving P, is the **Metropolis-Hastings algorithm**, which constructs a P-reversible transition kernel from a **proposal kernel** T. Assuming that $T(y \to \mathrm{d}x)P(\mathrm{d}y)$ and $T(x \to \mathrm{d}y)P(\mathrm{d}x)$ are equivalent (i.e., they are absolutely continuous with respect to each other) define

$$\alpha(x \to y) \equiv \min\left(1, \frac{T(y \to \mathrm{d}x)P(\mathrm{d}y)}{T(x \to \mathrm{d}y)P(\mathrm{d}x)}\right).$$

Then the transition kernel

$$\mathbb{P}(x \to \mathrm{d}y) \equiv \alpha(x \to y)T(x \to \mathrm{d}y) + \left(1 - \int \alpha(x \to z)T(x \to \mathrm{d}z)\right)\delta_x(\mathrm{d}y)$$

is P-reversible and thus preserves the target distribution [329]. Realisations of the Markov chain of $\mathbb{P}$ may be constructed as follows: given current x_n

Algorithm 1 Metropolis-Hastings algorithm, given current state x_n and proposal kernel T.

1: propose $x^* \sim T(x_n \to \cdot)$
2: accept x^* with probability $\alpha(x_n \to x^*)$, in which case set $x_{n+1} \equiv x^*$. Else $x_{n+1} \equiv x_n$.

An overview of sampling methods is provided in [70], which classifies them in four categories [329]: **(i)** methods generating a i.i.d. sample (e.g., rejection method), **(ii)** methods approximating sample paths of a diffusion (for example the overdamped Langevin dynamics), **(iii)** methods generating approximations of extended deterministic dynamical systems (such as the Nosé-Poincaré method), and finally **(iv)** method based on a P-preserving Markov chain. We will now review a sampling method of type **(iv)** that will be the core of this thesis: Hamiltonian Monte Carlo (HMC). In §2.1.1 we will recall the equations of motion of classical mechanics which are used to define the proposal kernel of HMC, and in §2.1.2 we introduce the standard HMC algorithm based on a leapfrog integrator of the mechanics for a target $P \propto e^{-V}\mathrm{d}q$. Other common choices of integrators which can sometimes outperform leapfrog are the subject of §2.1.3. In the appendix we provide a more detailed literature review of HMC. In §A.4.1 we elucidate the origins of the HMC algorithm, which arises from the need to sample from distributions on gauge groups for lattice quantum chromodynamics (QCD). We will see the method targets a distribution on the Lie group SU(3). Thus, despite what is commonly believed in the statistics community, HMC has always suited to sample distributions on manifolds, unlike

"Riemannian manifold HMC" which we will see samples distributions on $\mathbb{R}^n$![1] The equations of motion used in HMC on SU(3) were derived in [186] by explicitly constructing a symplectic structure from the Maurer Cartan frame. In §3.9 we will show this symplectic structure corresponds to a reduced mechanics consistent with the symmetries of the group. We also examine some of the most important "extensions" of HMC, including shadow HMC §A.4.2, stochastic gradient HMC in §A.4.3, HMC on Hilbert spaces §A.4.4, and discuss ergodicity in §A.4.5.

2.1.1 Mechanics

The construction of mechanics-based sampling algorithms relies on the interpretation of the log-density of the target as a potential energy, which defines a Newtonian mechanical system that is then used to generate proposals. Recall the trajectory $q(t) \in \mathbb{R}^3$ of a particle of mass $m > 0$ moving under the influence of a potential energy V satisfies Newton's equations[2]

$$m\ddot{q} = -\nabla_q V. \tag{2.2}$$

This can be restated in terms of Hamiltonian mechanics as saying the curve $t \mapsto (q(t), p(t)) \in \mathbb{R}^3 \times \mathbb{R}^3$ satisfies Hamilton's equations

$$\dot{q} = \nabla_p H, \qquad \dot{p} = -\nabla_q H \tag{2.3}$$

for $H(q,p) \equiv \frac{1}{2m}p^2 + V(q)$; and in term of Lagrangian mechanics by saying that $t \mapsto q(t)$ satisfies Hamilton's principle (of stationary action, §A.5.3) which leads to the Euler-Lagrange equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial v^i}\right)\Big|_{\dot{q}} = \frac{\partial \mathcal{L}}{\partial q^i} \tag{2.4}$$

for the Lagrangian $\mathcal{L}(q,v) \equiv \frac{1}{2}mv^2 - V(q)$. For this specific system, the above formulations of mechanics are equivalent, since the transformation from velocity to momenta $v \mapsto p = mv$ relates solutions of (2.4) to solutions of (2.3), but in general there is no relation between Hamiltonian and Lagrangian mechanics.[3] As a result of the measure-preserving properties of the mechanical flows both the Lagrangian and Hamiltonian formulations of mechanics are useful to construct the theory of HMC on manifolds as we shall see in §3, but in the rest of this introduction we will focus on the Hamiltonian framework.

[1] The reason for that name in [131] is to emphasize the statistical geometry induced by the Bayesian context which naturally generates a semi-Riemannian metric on sample space, see §5.1.4

[2] These assume the force field F is conservative. More generally we have $\frac{\mathrm{d}(m\dot{q})}{\mathrm{d}t} = F$.

[3] Indeed Lagrangian mechanics is based on a second-order equation which is only possible in velocity space $T\mathcal{M}$, while Hamiltonian mechanics is defined on momentum space $T^*\mathcal{M}$ as it requires a symplectic structure (see sec 3.5 [2]).

2.1.2 Hamiltonian Monte Carlo on Euclidean Space

The canonical measure on $\mathbb{R}^n$ is the Lebesgue measure $\mathrm{d}q$, so we assume our target distribution has the form $P \propto e^{-V}\mathrm{d}q$. The cotangent-lifted coordinates (q,p) form a Darboux chart of the canonical symplectic manifold $T^*\mathbb{R}^n \cong \mathbb{R}^n \times \mathbb{R}^n$, so the symplectic measure ω^n may be viewed as the product of the Lebesgue measure $\mathrm{d}q$ on $\mathbb{R}^n$, and the Lebesgue measure $\mathrm{d}p$ on the cotangent spaces, $\omega^n = \mathrm{d}q \wedge \mathrm{d}p$. To sample from P with HMC we need to lift P to a distribution $\propto e^{-H(q,p)}\mathrm{d}q\mathrm{d}p$ on $T^*\mathbb{R}^n$, associated with a Hamiltonian $H : T^*\mathbb{R}^n \to \mathbb{R}$, whose pushforward under the projection $(q,p) \mapsto q$ is P. The simplest way to construct H is to choose a Riemmanian metric $G : q \mapsto (G(q))_{ij}$ on $\mathbb{R}^n$ adapted to the target distribution, for example by fitting it to the Hessian of the target density (see [41, 236, 203] for a discussion on the choice of metric). We then define the following n-form[4] ζ

$$\zeta(q,p) = \frac{1}{\sqrt{(2\pi)^n \det G(q)}} \exp\left(-\tfrac{1}{2}p^\top G(q)^{-1}p\right) \mathrm{d}p|_q,$$

and the Hamiltonian

$$H(q,p) \equiv V(q) + \tfrac{1}{2}\log\left((2\pi)^n \det G(q)\right) + \tfrac{1}{2}p^\top G(q)^{-1}p.$$

With these definitions the lifted distribution

$$e^{-H(q,p)}\mathrm{d}q \wedge \mathrm{d}p = e^{-V(q)}\mathrm{d}q \wedge \zeta(q,p)$$

has the desired property that its marginal over q (integrating over the fibres) is indeed the target P. In the Hamiltonian H above we recognise the kinetic energy $T(q,p) \equiv \frac{1}{2}p^\top G(q)^{-1}p$ associated to the Riemannian metric G, whose Hamiltonian vector field is the geodesic flow, and a shifted potential energy $\tilde{V} \equiv V + \frac{1}{2}\log\left(\det G(q)\right)$ which is simply the density of P with respect to the Riemannian measure (see §A.5.1)

$$P \propto e^{-V}\mathrm{d}q = e^{-\tilde{V}+\frac{1}{2}\log(\det G(q))}\mathrm{d}q = e^{-\tilde{V}} \underbrace{\sqrt{\det G(q)}\mathrm{d}q}_{\text{Riemannian measure}}.$$

In fact we will see later that presence of the term $\frac{1}{2}\log\left((2\pi)^n \det G(q)\right)$ in the Hamiltonian is precisely to ensure we sample from $e^{-V}\mathrm{d}q$ rather than $e^{-V}\sqrt{\det G(q)}\mathrm{d}q$. When the geodesic flow Φ^T is tractable, the algorithm 2 based on the leapfrog integrator is sometimes called **Riemannian manifold HMC** or **geodesic Lagrangian Monte Carlo** (though no manifold

[4] This is indeed an n-form due the the transformation properties of the cotangent-lifted coordinates $p_j = P_k \frac{\partial Q^k}{\partial q^j}$ induced by a change of coordinates $Q \leftarrow q$ on $\mathbb{R}^n$, see definition 3.2.8 [163] (here P_k and p_k are the cotangent-lifted coordinate of Q and q respectively).

or Lagrangian is involved). If G is a constant matrix this simplifies to the standard HMC algorithm 3.

Algorithm 2 HMC Algorithm to generate N samples from $P \propto e^V \mathrm{d}q$ with geodesic integrator on Euclidean space, with time step δt, and $k \equiv \tau/\delta t$ where τ is the trajectory length of the integrator.

procedure EUCLIDEAN-RIEMANNIAN-HMC(N)
  for for $r \in [1, N]$ **do**
    $q_0 \equiv x_{r-1}$
    sample $v_0 \sim \mathcal{N}(0, G(q_0)^{-1})$
    $H_0 = V(q_0) + \frac{1}{2}\log\det G(q_0) + \frac{1}{2}v_0^\top G(q_0)v_0$
    $v_{\frac{1}{2}} = v_0 - \frac{1}{2}\delta t G^{-1}(q_0)\nabla_q \left(V + \frac{1}{2}\log\det G\right)|_{q_0}$
    $(q_1, v_1^*) \equiv \Phi_{\delta t}^T(q_0, v_{\frac{1}{2}})$
    for for $m \in [1, k-1]$ **do**
      $v_{m+\frac{1}{2}} \equiv v_m^* - \delta t G^{-1}(q_m)\nabla_q \left(V + \frac{1}{2}\log\det G\right)|_{q_m}$
      $(q_{m+1}, v_{m+1}^*) \equiv \Phi_{\delta t}^T(q_m, v_{m+\frac{1}{2}})$
    $v_k \equiv v_k^* - \frac{1}{2}\delta t G^{-1}(q_k)\nabla_q \left(V + \frac{1}{2}\log\det G\right)|_{q_k}$
    $H_f = V(q_k) + \frac{1}{2}\log\det G(q_k) + \frac{1}{2}v_k^\top G(q_k)v_k$
    $\delta H \equiv H_f - H_0$
    $U \equiv \mathrm{Uniform}[0, 1]$
    if $U \leq \exp(-\delta H)$ **then** $x_r \equiv q_k$
    else $x_r \equiv q_0$
  return $(x_r)_{r=1}^N$

2.1.3 Numerical Integration of Hamiltonian Flow

The most computationally demanding step of HMC algorithms is the numerical simulation of the Hamiltonian flow - it is thus crucial to choose carefully the numerical integrator by considering the trade-off between accuracy of the method (which leads to higher acceptance rate), its stability, and the computational effort required to implement it. While the most commonly used integrator is the leapfrog method, there has been a growing interest in considering more general (parametric) families of integrators and tune the parameters so as to outperform leapfrog for Hamiltonians of the form $H(q,p) = \frac{1}{2}p^\top G^{-1}p + V(q)$ on $\mathbb{R}^n \times \mathbb{R}^n$. The idea is to consider **palindromic multi-stage integrators**, specifically the **two-stage integrator** (recall Φ^H denotes the Hamiltonian flow of H and δt is the time-step)

$$\Psi_{\delta t} \equiv \Phi^V_{b\delta t} \circ \Phi^T_{\frac{1}{2}\delta t} \circ \Phi^V_{(1-2b)\delta t} \circ \Phi^T_{\frac{1}{2}\delta t} \circ \Phi^V_{b\delta t}$$

Algorithm 3 Standard HMC Algorithm to generate N samples from $P \propto e^V \mathrm{d}q$ with leapfrog integrator on Euclidean space, with time step δt, and with $k \equiv \tau/\delta t$ where τ is the trajectory length of the integrator.

procedure STANDARD-HMC(N)
  for for $r \in [1, N]$ **do**
    $q_0 \equiv x_{r-1}$
    sample $p_0 \sim \mathcal{N}(0, G)$
    $H_0 = V(q_0) + \frac{1}{2} p_0^\top G^{-1} p_0$
    $p_{\frac{1}{2}} = p_0 - \frac{1}{2}\delta t \partial_q V|_{q_0}$
    $q_{\delta t} = q_0 + \delta t G^{-1} p_{\frac{1}{2}\delta t}$
    for for $m \in [1, k-1]$ **do**
      $p_{m+\frac{1}{2}} = p_{m-\frac{1}{2}} - \delta t \partial_q V|_{q_0}$
      $q_{m+1} = q_m + \delta t G^{-1} p_{m+\frac{1}{2}}$
    $p_k = p_{k-\frac{1}{2}} - \frac{1}{2}\delta t \partial_q V|_{q_k}$
    $H_f = V(q_k) + \frac{1}{2} p_k^\top G^{-1} p_k$
    $\delta H \equiv H_f - H_0$
    $U \equiv \mathrm{Uniform}[0, 1]$
    if $U \leq \exp(-\delta H)$ **then** $x_r \equiv q_k$
    else $x_r \equiv q_0$
  return $(x_r)_{r=1}^N$

which contains leapfrog as a special case,[5] and **three-stage integrator**

$$\Psi_{\delta t} \equiv \Phi^V_{b\delta t} \circ \Phi^T_{a\delta t} \circ \Phi^V_{(\frac{1}{2}-b)\delta t} \circ \Phi^T_{(1-2)a\delta t} \circ \Phi^V_{(\frac{1}{2}-b)\delta t} \circ \Phi^T_{a\delta t} \circ \Phi^V_{b\delta t}.$$

For the purpose of sampling, these integrators should not be tuned to achieve the highest accuracy in the limit where $\delta t \to 0$ (as typically desired for numerical integrators), since in HMC we aim to make δt as large as possible in order to reduce the correlations between samples. Instead [301] proposed to tune the parameters by demanding that these preserve the energy during harmonic motion to order at least 4 (harmonic motions corresponds to a quadratic potential $V(q) = \frac{1}{2} q^\top q$, or equivalently a Gaussian target distribution $P(\mathrm{d}q) \propto e^{-\frac{1}{2} q^\top q} \mathrm{d}q$) and with time-step subject to the stability criteria that the numerical trajectories should remain bounded asymptotically. The performances of these **higher-order harmonic integrators** (HOH) in the context of HMC is analysed in 6 [301], where the authors show that by combining an HOH integrators with a RESPA splitting of the force into a fast and a slowly varying part (see [345]) one can use time steps twice as large as leapfrog with the same acceptance rate. A similar approach was implemented in [49] but where the tuning relies on optimising $\max_{\delta t \in (0, \overline{\delta t})} \varrho(\delta t)$ over the parameters, where the quantity $\varrho(\delta t)$ upper bounds the expected energy change during

[5] this follows by rewriting $\Phi^V_{(1-2b)\delta t}$ in $\Psi_{\delta t}$ as $\Phi^V_{(\frac{1}{2}-b)\delta t} \circ \Phi^V_{(\frac{1}{2}-b)\delta t}$ and setting $b = \frac{1}{4}$, which leads to a composition of two leapfrog steps with time-steps $\frac{1}{2}\delta t$: $\Phi^V_{\frac{1}{4}\delta t} \circ \Phi^T_{\frac{1}{2}\delta t} \circ \Phi^V_{\frac{1}{4}\delta t} \circ \Phi^V_{\frac{1}{4}\delta t} \circ \Phi^T_{\frac{1}{2}\delta t} \circ \Phi^V_{\frac{1}{4}\delta t}$

(numerical) harmonic motion, and $\overline{\delta t}$ is equal to the number of stages. For two-stages methods, an adaptive algorithm was proposed in [118] to suit the quantity $\overline{\delta t}$ to the time-step and the frequencies of the problem specified by the user. The algorithm optimises over the parameter b and outputs a two-stage integrator that shifts between the leapfrog method for large time-steps (because leapfrog has the largest stability region) and the optimal two-stage integrator of [258] for $\delta t \approx 0$. See also [68] for a comparison of palindromic 3-stage methods, and [54] for a recent survey of palindromic integrators for HMC. Note that as explained in [69, 90] given an integrator (therein leapfrog) it is easy to construct, through a recursive relation, symplectic reversible integrators with arbitrary high order of accuracy, see also [366] and Appendix C [85].

The tuning of the time-step as the dimension of the sample space tends to infinity is analysed in [39] for a product target measure $\otimes_{i=1}^{d} e^{-V(q)} \mathrm{d}q$ on $\mathbb{R}^{n \times d}$. It is shown that under suitable assumptions, for a volume-preserving, symmetric, and time-reversible integrator of order ν [6] with step-size $\delta t \equiv l d^{-\frac{1}{2\nu}}$, for a constant $l > 0$, the expected acceptance rate converges to a non-zero constant as $d \to \infty$. For the leapfrog integrator it is further argued that the optimal value of l (and thus δt) should give rise to an acceptance probability that is at least 0.651 (see theorem 4.1 therein for the optimality-criterion considered to reach this conclusion).

2.1.3.1 Numerical Integration with Multiple Time Scales

Sometimes the potential energy V can be decomposed as a sum $V \equiv V_{easy} + V_{hard}$ where $\hat{V}_{easy}$ requires far less computational power to evaluate than $\hat{V}_{hard}$, or varies more slowly. This happens for example in molecular systems where some of the forces acting on the system are long range and slowly varying, while other are short range and rapidly varying, or in lattice QCD where the gauge field self-interaction (the Wilson gauge action) is easier to compute than the pseudofermion term (see §A.4.1). In that case it is convenient to decompose the numerical integrator as [345, 315]

$$\Psi_{\delta t} \equiv \Phi^{V_{hard}}_{\frac{1}{2}\delta t} \circ \left(\Phi^{V_{easy}}_{\frac{1}{2}\delta t/k} \circ \Phi^{T}_{\delta t/k} \circ \Phi^{V_{easy}}_{\frac{1}{2}\delta t/k} \right)^{k} \circ \Phi^{V_{hard}}_{\frac{1}{2}\delta t}$$

for some $k \in \mathbb{N}$. Hence the more expensive flow is effectively computed over a longer time scale, the shorter time scale of the easier flow enables to improve the acceptance rate of Ψ without increasing its computational cost too much. Such a decomposition has been implemented early on in the context of lattice QCD [315], leading to substantial improvement over leapfrog integration (note for $k = 1$ we recover leapfrog). Effectively this procedure introduces a new parameter k, which in QCD can be tuned using Poisson brackets [187]. See also [316] for some

[6] Recall the numerical method $\Psi_{\delta t}$ is said to be **of order** ν with respect to the true trajectory $\Phi^{H}_{\delta t}$ if $\|\Psi_{\delta t} - \Phi^{H}_{\delta t}\| = O(\delta t^{\nu+1})$

applications in statistics.

2.2 Estimation of Probability Measures

The problem of **parameter inference** consists of estimating an element $\theta^* \in \Theta$ using a sequence of random functions (or estimators) $\hat{\theta}_n : \Omega \to \Theta$, with $\hat{\theta}_n$ determined by a set of observable random variables $\{X_1, \dots, X_n\}$ representing the available experimental data. For our purposes we are interested in approximating a target distribution P based on a sample using a statistical model $\{P_\theta : \theta \in \Theta\}$ (see §5.1.1), and thus we want to construct estimators $\hat{\theta}_n$ s.t., given a realisation ω of a $\otimes_{i=1}^n P$-distributed random function, the distribution P_{θ^*} which best approximates (in an appropriate sense) P within the statistical model is estimated to be $P_{\hat{\theta}_n(\omega)}$. Of course estimators will be of interest only if they possess theoretical properties that guarantees their usefulness. For example it is reasonable to expect the estimator $\hat{\theta} : \Omega \to \Theta$ to be **unbiased**, that is

$$\mathbb{E}[\hat{\theta}] = \theta^*.$$

If this relation does not necessarily hold we call the difference $\mathbb{E}[\hat{\theta}] - \theta^*$ the bias of the estimator. When we consider sequence of estimators $(\hat{\theta}_n)$ built from an increasing amount of data we may also want to ensure it satisfies some desirable asymptotic properties, such as asymptotic consistency. The sequence is said to be **weakly (strongly) consistent** if it converges in probability (almost surely) to θ^*

$$\text{weak consistency:} \quad \hat{\theta}_n \xrightarrow{p} \theta^*, \qquad \text{strong consistency:} \quad \hat{\theta}_n \xrightarrow{a.s.} \theta^*.$$

Moreover when $\Theta = \mathbb{R}^s$, a consistent sequence of estimators is **asymptotically normal** if

$$\sqrt{n}\left(\hat{\theta}_n - \theta^*\right) \xrightarrow{d} \mathcal{N}(0, V)$$

for some covariance matrix $V \in \mathbb{R}^{s\times s}$. For large n we have $\hat{\theta}_n \approx \frac{1}{\sqrt{n}}\mathcal{N}(0, V) + \theta^* = \mathcal{N}(\theta^*, V/n)$, so V/n (or sometimes V) is called the **asymptotic variance**. Asymptotic normality is a powerful guarantee that can be used for example to construct confidence intervals [277]. It also raises the question of how to construct estimators with the smallest possible asymptotic variance, or more generally the best asymptotic distribution. If $\{p_\theta \mu_\Theta : \theta \in \Theta\}$ is a parametrised statistical model with respect to some reference measure μ_Θ (see §5.1.2), the Cramér–Rao inequality shows that under appropriate regularity conditions for the model, if $\{X_1, \dots, X_l\}$ are independent $p_{\theta^*}\mu_\Theta$-distributed random variables, and $\hat{\theta}(X_1, \dots, X_l)$ is an unbiased estimator

of θ^*, then the covariance of $\hat{\theta}$ is bounded below by the inverse Fisher information matrix $I(\theta^*)$ (i.e., their difference is positive semidefinite)

$$\mathrm{Cov}(\hat{\theta}) \geq I(\theta^*)^{-1}.$$

This result is only valid for a single estimator. Chapter 8 [348] discusses an "asymptotic Cramér–Rao", and shows that under some condition the asymptotic distribution $\mathcal{N}(0, I(\theta^*)^{-1})$ for a sequence of estimators may be considered optimal, although this is not completely true. Nonetheless asymptotically normal sequences satisfying $\sqrt{n}\left(\hat{\theta}_n - \theta^*\right) \xrightarrow{d} \mathcal{N}(0, I(\theta^*)^{-1})$ are usually referred to as **asymptotically efficient/optimal**.

A particular instance of distribution inference which plays a central role in statistics is the problem of **density estimation**. In the Fisher–Wald setting one considers a family of probability densities $\{p_\theta : \theta \in \Theta\}$ containing the target measure $P = p_{\theta^*}\mu_\Theta$. The aim is to minimise the risk functional

$$R[p_\theta] \equiv -\int \log p_\theta(x) P(\mathrm{d}x),$$

since if a function q minimises R, then $q = p_{\theta^*}$ μ_Θ-almost everywhere, and moreover by the Bretagnolle–Huber inequality $d(p_\theta, p_{\theta^*}) \leq 2\sqrt{1 - e^{R[p_{\theta^*}] - R[p_\theta]}}$, where d is the $L^1(\mu_\Theta)$ metric (see 1.6 [350]). Given random variables $\{X_1, \ldots, X_n\}$, the estimator $\hat{\theta}_n$ is defined using the principle of empirical risk minimisation in which we formally replace the measure P in the risk with the empirical measure $\frac{1}{n}\sum_i^n \delta_{X_i}$ associated to the experiment

$$\hat{\theta}_n \equiv -\mathrm{argmin}_{\theta\in\Theta} \frac{1}{n} \sum_{i=1}^n \log p_\theta(X_i).$$

This estimator is also known as the **maximum likelihood estimator** (MLE) when written as a maximiser

$$\hat{\theta}_n \equiv \mathrm{argmax}_{\theta\in\Theta} \frac{1}{n} \sum_{i=1}^n \log p_\theta(X_i) = \mathrm{argmax}_{\theta\in\Theta} \log p_\theta^{\otimes^n}(X_1, \ldots, X_n), \tag{2.5}$$

where $p_\theta^{\otimes^n}$ on the RHS is the density of the joint distribution $\otimes_{i=1}^n P_\theta$ of the data. When Θ is Euclidean (with Cartesian coordinates), the optimisation (2.5) is commonly tackled through (Euclidean) gradient ascent (see §5.2). When $p_\theta \equiv \frac{1}{Z_\theta} e^{-E_\theta}$, since

$$\begin{aligned}\partial_\theta \frac{1}{n} \sum_{i=1}^n \log p_\theta(X_i) &= -\frac{1}{n} \sum_{i=1}^n \partial_\theta E_\theta(X_i) - \partial_\theta \log Z_\theta = -\frac{1}{n} \sum_{i=1}^n \partial_\theta E_\theta(X_i) - \partial_\theta \log \int e^{-E_\theta(x)} \mathrm{d}x \\ &= -\frac{1}{n} \sum_{i=1}^n \partial_\theta E_\theta(X_i) + \frac{\int e^{-E_\theta(x)} \partial_\theta E_\theta(x) \mathrm{d}x}{\int e^{-E_\theta(y)} \mathrm{d}y}\end{aligned}$$

$$= -\frac{1}{n}\sum_{i=1}^{n} \partial_\theta E_\theta(X_i) + \int \partial_\theta E_\theta \mathrm{d}P_\theta,$$

a step of gradient ascent with $\eta > 0$ reads

$$\hat{\theta} \leftarrow \hat{\theta} + \eta \left(-\frac{1}{n}\sum_{i=1}^{n} \partial_\theta E_\theta(X_i) + \int \partial_\theta E_\theta \mathrm{d}P_\theta \right). \tag{2.6}$$

Another class of estimators was proposed in [292] where the estimated density is defined by in terms of a sequence of bandwidth parameters $n \mapsto h(n) > 0$ and function k

$$p_{\hat{\theta}_n}(x) \equiv \frac{1}{nh(n)} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h(n)}\right).$$

Conditions are provided for these estimators to have desirable asymptotic properties. When k is a kernel function, these are also referred to as the **kernel density estimators** (KDE). The article [292] further analysed their application to the problem of mode estimation, where the mode $\alpha \equiv \mathrm{argmax}_x p_{\theta^*}(x)$ of the target density (assumed unimodal) is estimated from the "sample modes" $\hat{\alpha}_n \equiv \mathrm{argmax}_x p_{\hat{\theta}_n}(x)$. One can consider more general "adaptive" kernel estimators where the bandwidth h is allowed to depend on x or on the data X_i [339], which may for example be used to detect rare events (outliers/anomalies) such as fraudulent transactions [206]. Estimators can also be built for distributions over functions, such as those arising from stochastic processes generated by diffusions [200]. For example [15] considers the problem of estimating the drift and diffusion terms by minimising the variational free energy between the posterior distribution over sample path and a family of Gaussian processes. Below we discuss some important classes of estimators, in particular the minimum Hyvärinen score estimator which is commonly used for unnormalised differentiable statistical models, and will be generalised in §6 to improve its robustness. Additional estimators for likelihood-free inference based on generative models are reviewed in §A.4.

2.2.1 M, Z, and Wasserstein Distance Estimators

Given a family of functions $m_\theta : \mathcal{M} \to \mathbb{R} \bigcup \{\infty\}$ and i.i.d. random functions $X_i : \Omega \to \mathcal{M}$, the estimators

$$\hat{\theta}_n \equiv \mathrm{argmax}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i) \tag{2.7}$$

are called **M-estimators**. The MLE defined in (2.5) is the M-estimator associated to the log-likelihood functions $m_\theta \equiv \log p_\theta$. When Θ is an open subset of Euclidean space and $\theta \mapsto m_\theta$

is differentiable, by the first optimality condition a maxima θ^* will satisfy

$$\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta m_\theta(X_i)\big|_{\theta^*}=0, \tag{2.8}$$

and in practice solutions to (2.7) are often found by solving (2.8). However since these equations are not equivalent, it is better to refer to solution to (2.8) as **Z-estimators** (see 5 [348]). The parameter space Θ does not need to be finite dimensional. For example we can consider the exponential family $P_\theta \equiv \frac{1}{Z(\theta)}e^{\theta(x)}\mathrm{d}x$ where θ belongs to an appropriate subset of a reproducing kernel Hilbert space (see §A.4.9). The MLE then reads

$$\hat{\theta}_n = \mathrm{argmax}_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\theta(X_i)-\log Z(\theta),$$

and the corresponding Z-estimator satisfies (see (1.7) [123])

$$\int k(x,\cdot)\mathrm{d}P_\theta = \frac{1}{n}\sum_{i=1}^{n}k(X_i,\cdot),$$

i.e., the mean embedding of the estimated distribution is the mean embedding of the empirical measure. To obtain tractable estimators [123, 1.3.3] discusses possible approaches to regularise the objective, in particular the approximation of Θ by an increasing sequence of finite dimensional subspaces.

An important mechanism to generate useful M-estimators relies on the concept of proper scoring rules. A **scoring rule** is a function $S:\mathcal{M}\times\mathcal{P}(\mathcal{M})\to\mathbb{R}$ which measures the discrepancy between a realisation of an $\mathcal{M}$-valued random variable and a probability measure on $\mathcal{M}$ (here $\mathcal{P}(\mathcal{M})$ is the set of probability measures on $\mathcal{M}$): the bigger $S(x,Q)$ is, the less likely it is that x was generated by a Q-distributed random variable. It is a **(strictly) proper scoring rule** if the map $Q\mapsto\int S(x,Q)P(\mathrm{d}x)$ is (uniquely) minimised at P. A proper scoring rule, defines a **minimum score estimator** by

$$\hat{\theta}_n \equiv \operatorname*{argmin}_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}S(X_i,P_\theta).$$

For example, given a convex differentiable function $\psi:\mathbb{R}_{\geq 0}\to\mathbb{R}$, assuming $Q(\mathrm{d}z)=q(z)\mathrm{d}z$, the **Bregman scoring rule** is defined as

$$S(x,Q)\equiv -\psi'(q(x))-\int\left(\psi(q(z))-q(z)\psi'(q(z))\right)\mathrm{d}z.$$

It is proper, strictly proper if ψ is strictly convex, and recovers MLE for $\psi(z)=z\ln z$. Of

particular importance are **homogeneous** scoring rules which are invariant under scaling in their second entry, and thus may be evaluated without knowledge the normalisation constant (see 6 [291] for a detailed study). The most famous homogeneous scoring rule is the **Hyvärinen score** on $\mathcal{M} = \mathbb{R}^n$, when $q > 0$, associated to the score matching statistical divergence §5.1.3

$$S(x, Q) \equiv \tfrac{1}{2} \left\| \nabla \log q \right\|^2 + \Delta \log q. \tag{2.9}$$

Another important class of estimators originate from optimal transport. Therein one aims to solve the **Monge–Kantorovich transportation problem**, that is to find the transfer plan $\Pi_{P \leftrightarrow Q}$ between two measures that minimises a given **cost** function $c : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_+$ representing the work required to transport probability mass [352],

$$\inf_{\Pi_{P \leftrightarrow Q}} \int_{\mathcal{M} \times \mathcal{M}} c(x, y) \Pi_{P \leftrightarrow Q}(\mathrm{d}x, \mathrm{d}y).$$

If we fix P, and consider a statistical model parametrised by a metric space Θ, we can define **Kantorovich risk**

$$R[P_\theta] \equiv \inf_{\Pi_{P \leftrightarrow P_\theta}} \int_{\mathcal{M} \times \mathcal{M}} c(x, y) \Pi_{P \leftrightarrow P_\theta}(\mathrm{d}x, \mathrm{d}y),$$

and given n independent P-distributed random variables (X_i) the **minimum Kantorovich estimator** (MKE) $\hat{\theta}_n$ is defined by

$$\hat{\theta}_n \equiv \operatorname*{argmin}_{\theta \in \Theta} \inf_{\Pi_{\frac{1}{n}\sum \delta_{X_i} \leftrightarrow P_\theta}} \int_{\mathcal{M} \times \mathcal{M}} c(x, y) \Pi_{\frac{1}{n}\sum \delta_{X_i} \leftrightarrow P_\theta}(\mathrm{d}x, \mathrm{d}y).$$

Conditions for the existence, measurability, and consistency of MKEs are given in [30]. When $(\mathcal{M}, d)$ is a Polish metric space, $p \in [1, \infty)$ the **p-Wasserstein distance** between any two distributions P, Q on $\mathcal{M}$ is

$$W_p(P, Q) \equiv \left(\inf_{\Pi_{P \leftrightarrow Q}} \int_{\mathcal{M} \times \mathcal{M}} d(x, y)^p \Pi_{P \leftrightarrow Q}(\mathrm{d}x, \mathrm{d}y) \right)^{\frac{1}{p}}. \tag{2.10}$$

It defines a metric on the **Wasserstein space** of Borel distributions Q with finite p-moment, $\int_{\mathcal{M}} d(x_0, x)^p \mathrm{d}Q(x) < \infty$, for some arbitrary $x_0 \in \mathcal{M}$ (else the Wasserstein distance can be infinite). When $\mathcal{M}$ is compact, the Wasserstein distance metrises the narrow topology (see §A.2)

$$Q_n \to P \text{ narrowly} \qquad \text{iff} \qquad W_p(Q_n, P) \to 0;$$

while when $\mathcal{M}$ is not compact, one must add a convergence of moment condition, see theorem 6.8 [352]. For $p = 1$ we can equivalently use the dual formulation for distributions on the

Wasserstein space

$$W_1(P,Q) = \sup_{\|\psi\|_{\text{Lip}} \leq 1} \left(\int \psi \mathrm{d}P - \int \psi \mathrm{d}Q \right)$$

where $\|\psi\|_{\text{Lip}} \equiv \sup \left\{ \frac{|\psi(x)-\psi(y)|}{d(x,y)} : x \neq y \right\}$ for $\psi \in C(\mathcal{M})$. On Euclidean space $\mathcal{M} \equiv \mathbb{R}^d$ with Euclidean metric, the Lipschitz constraint can be replaced by a gradient constraint (6.2 [298])

$$W_1(P,Q) = \sup_{\|\nabla\psi\|_\infty < 1} \left(\int \psi \mathrm{d}P - \int \psi \mathrm{d}Q \right).$$

A useful feature of the Wasserstein distance it is well-defined on discrete distributions, in particular $W_p(\delta_x, \delta_y) = d(x,y)$. If we set $\mathcal{M} = \{x_1, \ldots, x_n\}$ to be a discrete set with n elements, the distributions on $\mathcal{M}$ are elements of the (probability) simplex (also called histograms), $P \equiv (p_1, \ldots, p_n) \in \mathbb{R}^n_{\geq 0}$ s.t., $\sum_i p_i = 1$, the metric d defines a distance matrix $D \in \mathbb{R}^{n \times n}_{\geq 0}$ with components $D_{ij} = d(x_i, x_j)$, and a transport plan $\Pi_{P \leftrightarrow Q}$ is a nonnegative matrix $\boldsymbol{P}_{P \leftrightarrow Q} \equiv \boldsymbol{P} \in \mathbb{R}^{n \times n}_{\geq 0}$ satisfying $\sum_j \boldsymbol{P}_{ij} = p_i$, $\sum_i \boldsymbol{P}_{ij} = q_j$. The 1-Wasserstein distance is then given by the linear program (see 2.3-2.4, 3.1 [298])

$$W_1(P,Q) = \min_{\boldsymbol{P}_{P \leftrightarrow Q}} \boldsymbol{P}^\top D. \tag{2.11}$$

The **minimum Wasserstein distance** estimators are obtained by replacing P and P_θ by their empirical measures

$$\hat{\theta}_n \equiv \text{argmin}_{\theta \in \Theta} W_1(P_n, P_\theta), \qquad \hat{\theta}_{n,m} \equiv \text{argmin}_{\theta \in \Theta} W_1\left(P_n, P_{\theta,m}\right).$$

Their existence, measurability, consistency and asymptotic distribution is studied in [38]. In §A.4.8 we discuss the practical implementation of Wasserstein estimators via Sinkhorn statistical divergences, which are the main competitors to the minimum MMD estimators we proposed in [62] for inference in likelihood-free models (see §A.8).

Chapter 3

Hamiltonian Monte Carlo on Manifolds

The aim of this chapter is to develop a generalised and rigorous theory of Hamiltonian-based Monte Carlo samplers in a harmonised mathematical formalism, the language of geometric mechanics, discuss the relations between several variants of HMC, correct some common mistakes in the implementation of HMC on manifolds, show the importance of symmetries in the efficient implementation of the numerical trajectories, and in particular use reduction by symmetry to implement HMC on geodesic orbit manifolds by lifting the scheme to the acting Lie Group. Moreover, understanding the geometric theory on which HMC is built will allow us to generalise it in §4, and obtain a complete theory of MCMC methods generated by "brackets" (a generalisation of the non-degenerate Poisson structures on which HMC is based). While in Euclidean space we had a canonical choice of reference measure, the Lebesgue measure, this is no longer the case on a general manifold so we will discuss choices of reference measures in §3.1, as well as the geodesic flow which plays a central role in many HMC algorithms. Given an appropriate reference measure we can parametrise target distributions through their densities. We saw in §2.1.1 that HMC on $\mathbb{R}^n$ is based on defining a potential energy from the density of the target distribution, and uses Newtonian mechanics to propose samples. On manifolds we can follow the same idea, and we will discuss in §3.2 the various formulations of mechanics and their measure-preserving properties. In particular we will see that, essentially, the equation of motion on manifolds may be thought as simply replacing the flat Newtonian system with its Riemannian (i.e., curved) analogue (see §3.2.3)

$$\underbrace{m\ddot{\gamma} = -\partial V}_{\text{Flat Newton}} \quad \longrightarrow \quad \underbrace{\frac{\nabla\dot{\gamma}}{\mathrm{d}t} = -\nabla V}_{\text{Riemannian Newton}} ,$$

and as we shall see the Riemannian Newton flow preserves the "canonical distribution" $\propto e^{-\frac{1}{2}\|\cdot\|^2 - V}\omega_\flat^n$, whose pushforward under the quotient projection $T\mathcal{M} \to \mathcal{M}$ is $\propto e^{-V}\text{vol}$, where vol is the Riemannian measure and $\omega_\flat$ the "Riemannian" symplectic measure. This allows us to transport samples of the canonical distribution to samples of the target. In §3.3 we will explain how to generally lift the target measure on $\mathcal{M}$ to a symplectic fibre bundle $(\mathcal{F}, \omega)$ over $\mathcal{M}$ in a measure-preserving manner, so that we can then use the mechanics defined by ω to generate samples from the target. In §3.4 we shall then define the general Hamiltonian Monte Carlo algorithm, which combines a mechanical proposal, a Monte Carlo accept/reject step, and a heat bath, and discuss in full generality several standard implementations.

The theory of integrators suitable for HMC is the content of §3.5, where we will review geodesic integrators (the Hamiltonian perspective), and variational integrators (the Lagrangian perspective) obtained using the discrete variational principle, which are the ones of interest for sampling distributions on embedded manifolds when the geodesic flow is intractable. In theory, HMC only requires the integrator to be time-reversible, time-symmetric, and volume preserving (see §3.4.1), yet it is usually better to use integrators which are also symplectic. The general theory of symplectic maps (based on generating functions) is reviewed in §3.5.3. Combined with Hamilton-Jacobi PDE it implies the existence of a Shadow Hamiltonian, thus explaining the surprisingly high acceptance rate observed when implementing HMC with symplectic integrators (see also §A.5.6). Shadow Hamiltonians will also be important in the next chapter when we consider non-symplectic Poisson systems §4.

We will then discuss the implementation of HMC on manifolds for various general settings: embedded manifolds, submersed manifold, shape spaces and Lie groups. In §3.6 we will focus on embedded manifolds (not necessarily in $\mathbb{R}^n$). We will generalise and correct the algorithm proposed in [66], which has caused a lot of confusion (as well as the claims in [160]), and moreover show that both [161] and [186] are special cases of this general setting. In §3.7 we will consider the setting in which there is a Riemannian submersion onto $\mathcal{M}$. Unlike Riemannian embeddings (where the Riemannian structure of the embedded manifold is induced by the one of the ambient space) Riemannian submersions have powerful geodesic/gradient-preserving properties which simplify the implementation of HMC. Riemannian submersions arise naturally in the context of shape spaces (discussed in §3.8), and we will develop the general theory of reduction by symmetry needed to construct the mechanics on a shape space $\mathcal{M} \equiv \mathcal{E}/\mathcal{K}$ from a mechanics on $\mathcal{E}$. Importantly, for any geodesic orbit manifold $\mathcal{M} \equiv \mathcal{G}/\mathcal{K}$, which include hyperspheres, hyperbolic spaces, Stiefel manifolds, Grassman manifolds, and the space of positive definite matrices, we can implement HMC entirely on $\mathcal{G}$, and thus take advantage of the properties of $\mathcal{G}$ (such as $T^*\mathcal{G} \cong \mathcal{G} \times \mathfrak{g}^*$). In that case, the practical implementation of HMC via geodesic integrators may in fact be seen as combining a Riemannian submersion $\mathcal{G} \to \mathcal{G}/\mathcal{K}$ and a totally geodesic embedding $\mathcal{G} \to \text{GL}(n)$. Finally, since mechanics on Lie groups plays

a central role in HMC on manifolds, we will consider it carefully in §3.9. Specifically, we will generalise the original HMC on compact semisimple matrix groups of [186] by explaning its relation with the Euler–Poincaré formalism in geometric mechanics, and deriving the solutions to Euler–Arnold equation for various systems with symmetries, including $\mathrm{GL}(n)$ and naturally reductive manifolds.

We recall the table 1.1 for a summary of the notation that will be used in this chapter. We will assume standard notions from differential geometry (manifolds, differential forms...) and provide an intuitive description of these objects in §A.3.

3.1 Measures and Geodesics on Manifolds

In HMC we need to choose a reference on $\mathcal{M}$ in order to expand the target measure in terms of a density to construct its potential energy. On $\mathbb{R}^n$ the natural reference measure is the Lebesgue measure, but there is no such natural choice on a general manifold $\mathcal{M}$. In various papers the density and associated potential energy of the target distribution is defined in a local coordinate system as $P(\mathrm{d}q) = e^{-V}\mathrm{d}q$ (e.g. [66, 202]). However this does not give rise to a well-defined potential energy, since under another coordinate system y^i we have $P(\mathrm{d}y) = e^{-V}|\det(\frac{\partial q^i}{\partial y^j})|\mathrm{d}y$, which defines a new potential energy $V' \equiv V - \log|\det(\frac{\partial q^i}{\partial y^j})|$ (see also 3 [296]). In order to obtain a well-defined density we need to choose a well-defined reference measure on $\mathcal{M}$. If we equip $\mathcal{M}$ with a Riemannian metric it acquires an associated Radon measure as we remind in §A.5.1. However this construction depends on the arbitrary choice of metric. Fortunately many manifolds appearing in statistics and physics are Lie groups or quotient manifolds. As we will see in §3.1.1 for many Lie groups there is a canonical choice, the Haar measure, which generalises the Lebesgue measure. This is still true for some Lie group quotients $\mathcal{G}/\mathcal{K}$, for example when $\mathcal{K}$ is compact, where through the action of $\mathcal{G}$ we can define the canonical/uniform measure §3.1.2. As the implementation of HMC often involves computing geodesics we will discuss in §3.1.3 manifolds in which these may be found, the geodesic orbit manifolds. In the appendix §A.5.2 we also mention first-order approximations known as retractions that may be useful to implement HMC when the geodesic flow has a prohibitive computational cost.

3.1.1 Haar Measures

A **left Haar measure** μ on a Lie group $\mathcal{M} \equiv \mathcal{G}$ is a non-zero Borel measure that is invariant under the left-translations L_g for any $g \in \mathcal{G}$,

$$\int f \circ L_g \mathrm{d}\mu = \int f \mathrm{d}\mu$$

for any $f \in C_c(\mathcal{G})$. Replacing L_g with R_g defines right Haar measures. Left Haar measures are proportional to each other, they give positive measure to non-empty open set, and measure zero to lower-dimensional subsets of $\mathcal{G}$. We can construct the left Haar measure using the left-invariant volume form $\Theta \equiv \theta^1 \wedge \cdots \wedge \theta^n$ for a basis $\theta^i \in \mathfrak{g}_L^*$ of left-invariant 1-forms, since the associated positive measure μ defined by Riesz-representation, $\int f \mathrm{d}\mu = \int f\Theta$, is left-invariant $\int f \circ L_g \mathrm{d}\mu = \int L_g^* f\Theta = \int L_g^*(f\Theta) = \int f\Theta = \int f \mathrm{d}\mu$, where we used theorem 2.6.7 [2]. In particular when sampling with HMC on a Lie group $\mathcal{G}$ using a geodesic integrator (see §3.5.1) associated to a left-invariant metric $G_{ij}\theta^i \otimes \theta^j$ (where G_{ij} are the components of an inner product matrix), the reference measure on $\mathcal{G}$ should be set to be the left Haar measure, since it is necessarily proportional to the Riemannian volume form of the left-invariant metric (otherwise a correction term should be included in the Hamiltonian). For numerous groups, including compact, abelian and semisimple Lie groups, the left Haar measure is also right-invariant, and is simply referred to as the (invariant) Haar measure (hence one may use the geodesic flow of a right-invariant metric).

For example, on $\mathbb{R}^n$ viewed as the abelian additive group $(\mathbb{R}^n, +)$, the Haar measure is the Lebesgue measure, which is thus the unique (up to a constant) translation-invariant measure. Similarly, the Haar measure on the circle S^1 is the Lebesgue measure. On the scale group $\mathbb{R}_{>0}$ with multiplication, the Haar measure is $\mathrm{d}x/x$. The group of invertible matrices $\mathrm{GL}(n)$ is not compact/abelian/semisimple but still has a Haar measure given by $|\det x|^{-n}\mathrm{d}x$ (see section 6 chapter VIII [191]). A simple calculation shows its left-invariance:

$$\begin{aligned}\int f(gx)|\det x|^{-n}\mathrm{d}x &= \int f(y)|\det g^{-1}y|^{-n}|\det \partial L_{g^{-1}}|\mathrm{d}y \\ &= \int f(y)|\det g|^n|\det y|^{-n}|\det g^{-1}|^n\mathrm{d}y = \int f(y)|\det y|^{-n}|\mathrm{d}y\end{aligned}$$

where we have used $\det \partial L_{g^{-1}} = \det L_{g^{-1}} = (\det g^{-1})^n$ (so $\mathrm{d}x = |\det g^{-1}|^n\mathrm{d}y$). An analogous derivation proves the measure is also right-invariant.

3.1.2 Uniform Measures

For Lie groups the canonical/uniform measure is the Haar measure (whenever it exists), that is the measure invariant under left and right translations. Similarly for homogeneous manifolds $\mathcal{M} \equiv \mathcal{G}/\mathcal{K}$ [1] the **uniform measure** is the measure invariant under the natural (left) action of $\mathcal{G}$ on $\mathcal{G}/\mathcal{K}$, denoted $\Phi_g : h\mathcal{K} \mapsto gh\mathcal{K}$ (see 3 [297] and 4 [295]). While a uniform measure does not exist in general, sufficient conditions for its existence can be found in [323], and, in particular, when $\mathcal{K}$ is compact there exists a $\mathcal{G}$-invariant Riemannian metric on $\mathcal{G}/\mathcal{K}$ whose Riemannian measure is the uniform measure. In general, when a uniform measure exists it is unique up to a constant (since the action of $\mathcal{G}$ on $\mathcal{G}/\mathcal{K}$ is transitive), and when $\mathcal{G}/\mathcal{K}$ is compact it can be normalised to define the uniform probability distribution.

An important class of homogeneous manifolds are the **reductive** ones. This means that there exists a subspace $\mathfrak{p} \subset \mathfrak{g}$ with

$$\mathfrak{g} = \mathfrak{p} \oplus \mathfrak{k}, \qquad \mathrm{Ad}_k\, \mathfrak{p} \subset \mathfrak{p}, \quad \forall k \in \mathcal{K},$$

where $\mathrm{Ad} : \mathcal{G} \to \mathrm{GL}(\mathfrak{g})$ is the adjoint representation. We further assume that $\mathcal{M}$ is equipped with a $\mathcal{G}$-invariant metric, which, as mentioned above, can always be found when $\mathcal{K}$ is compact (proposition 22.22 [125]). Such $\mathcal{G}$-invariant metrics are in 1-to-1 correspondence with $\mathrm{Ad}_{\mathcal{G}}\, \mathcal{K} \equiv \mathrm{Ad}\,|_{\mathcal{K}}$-invariant inner products on $\mathfrak{p}$, which is the generalisation of the correspondence between inner products on $\mathfrak{g}$ and left-invariant metrics on $\mathcal{G}$. Moreover, in that case we can construct a left-$\mathcal{G}$-invariant and right-$\mathcal{K}$-invariant metric on $\mathcal{G}$ s.t. $\mathfrak{p} = \mathfrak{k}^{\perp}$, and the quotient projection $\tau : \mathcal{G} \to \mathcal{M}$ is a Riemannian submersion (see proposition 22.23 [125]). In particular the isometry $\mathfrak{p} \to T_{g\mathcal{K}}(\mathcal{G}/\mathcal{K})$ is given by $X \mapsto \partial_{\mathcal{K}}\Phi_g \circ \partial_1\tau(X) = (\mathrm{Ad}_g\, X)_{\mathcal{M}}(g\mathcal{K})$, which shows that the adjoint action Ad_g adapts the complementary subspace $\mathfrak{p}$ to the point of $g\mathcal{K}$ interest ($\mathrm{Ad}_g\, \mathfrak{k}$ is the Lie algebra of the isotropy group at $g\mathcal{K}$).[2] This fact will be particularly important when we implement on $\mathcal{G}$ an HMC scheme to sample from $\mathcal{G}/\mathcal{K}$, since we can refresh the momentum on the single tangent space $\mathfrak{p}$, and moreover never need to actually compute the adjoint action because we are only interested in the configuration samples (so we average over momentum before projecting). We finally mention that there is a simple relation between the uniform measure on $\mathcal{G}/\mathcal{K}$ and the left Haar measures on $\mathcal{G}$ and $\mathcal{K}$: up to a constant, the pull-back of the uniform (Riemannian) volume form $\mathrm{vol}_{\mathcal{G}/\mathcal{K}}$ is precisely (see 10 [323])

$$\theta^1 \wedge \cdots \wedge \theta^n = \tau^* \mathrm{vol}_{\mathcal{G}/\mathcal{K}},$$

[1] Note we always assume $\mathcal{K}$ is a closed subgroup of $\mathcal{G}$, so that $\mathcal{G}/\mathcal{K}$ is a manifold, whose elements are the cosets $g\mathcal{K}$ for $g \in \mathcal{G}$. $\mathcal{K}$ is the **isotropy group** which fixes the origin $1\mathcal{K} \in \mathcal{M}$.

[2] Here $\Phi_g : \mathcal{M} \to \mathcal{M}$ denotes the action of g, and $\xi_{\mathcal{M}}$ is the action vector field on $\mathcal{M}$ generated by the element of the Lie algebra ξ (see (3.17))

where $(\theta^i)_{i=1}^{i=n}$ is a basis of $\mathfrak{p}_L^*$ (the left-invariant 1-forms generated by $\mathfrak{p}^*$), so the left Haar measure of $\mathcal{G}$ has (up to a constant) the decomposition

$$\Theta_{\mathcal{G}} = \tau^* \mathrm{vol}_{\mathcal{G}/\mathcal{K}} \wedge \Theta_{\mathcal{K}},$$

which is useful in the context of fibre integration. A local instance of this decomposition consists of using the local Hopf-coordinates SO(3), which then decomposes as the product $\mathrm{SO}(3) \cong S^2 \times S^1$, so we can parametrise it with (θ, ϕ, ψ) where $\theta \in [0, \pi], \phi \in [0, 2\pi)$ are the usual coordinates on the sphere, and $\psi \in [0, 2\pi)$ is the coordinate on the circle. The Haar measure is $\sin\theta \mathrm{d}\theta \mathrm{d}\phi \mathrm{d}\psi$ [365], where we recognise the uniform spherical measure $\sin\theta \mathrm{d}\theta d\phi$ on S^2, and the uniform measure $\mathrm{d}\psi$ on the circle S^1.

3.1.3 Geodesic Flow

To sample from a distribution using HMC we need to compute the Hamiltonian vector field of the kinetic energy which is the geodesic flow of the Riemannian metric. Finding the geodesics is usually a difficult problem, so it is important to choose a Riemmanian metric which makes the geodesic flow tractable, as this simplifies considerably the implementation of HMC. The natural way to simplify the geodesics equations is to restrict to Riemannian metrics with symmetries. For example on a Lie group, the geodesics equations of left-invariant metrics reduce to two first order ODEs, the Euler-Arnold and reconstruction equations, rather than the usual second order ODE. If the Riemannian metric is further bi-invariant (that is both left and right invariant), the geodesics are then given by the Lie group exponential as discussed below. Bi-invariance is precisely the reason that the geodesic flow on $\mathbb{R}^n$ of the Euclidean metric is $t : (q, v) \mapsto q+tv$, and that for some matrix groups such as SU(3) it is given by a matrix exponentiation $(q, v) \mapsto qe^{tv}$.

As we shall now see, geodesics on $\mathcal{G}/\mathcal{K}$ which are related to Lie group exponentials by projections are called homogeneous geodesics, and for a class of manifolds known as **geodesic orbit manifolds** it is possible to choose a Riemannian metric whose geodesics are all homogeneous. In particular for naturally reductive manifolds these homogeneous geodesics have a simple description. When the geodesics do not have a closed-form expression, one may use a constrained numerical method to compute them (see appendix [211]) or a retraction §A.5.2.

3.1.3.1 Naturally Reductive Manifolds

A reductive homogeneous manifold $\mathcal{G}/\mathcal{K}$ with $\mathcal{G}$-invariant metric is said to be naturally reductive if the associated inner product on $\mathfrak{g}$ (recall §3.1.2) satisfies ($\cdot|_{\mathfrak{p}}$ denotes the $\mathfrak{p}$-component)

$$\langle [X,Y]_{\mathfrak{p}}, Z\rangle_{\mathfrak{g}} + \langle Y, [X,Z]_{\mathfrak{p}}\rangle_{\mathfrak{g}} = 0, \qquad \forall X,Y,Z \in \mathfrak{p}.$$

This is a natural generalisation of the notion of bi-invariant metric on Lie groups since for $\mathcal{K} = \{1\}$ the statement reduces to demanding that $\mathcal{G} \cong \mathcal{G}/\{1\}$ has an $\operatorname{ad}\mathfrak{g}$-invariant inner product (recall $\operatorname{ad} \equiv \partial_1 \operatorname{Ad}$), and for connected Lie groups, $\operatorname{ad}\mathfrak{g}$-invariant inner products are equivalent to bi-invariant Riemannian metrics (see §3.9.2). In that case it follows directly from Euler–Arnold equations that the geodesic through the point $g \in \mathcal{G}$ with initial velocity $X \in \mathfrak{g}$ is given by the curve $t \mapsto ge^{tX}$ (see §3.9.2.1). For a general naturally reductive manifold $\mathcal{G}/\mathcal{K}$, the geodesic through the point $g\mathcal{K}$ with initial velocity $v_{g\mathcal{K}}$ is given by (proposition 22.28 [125])

$$t \mapsto \tau\left(ge^{tX}\right) = \Phi_t^T\left(\tau(g), \left(\operatorname{Ad}_g X\right)_{\mathcal{M}}(\tau(g))\right)$$

where Φ^T is the geodesic flow, and $X \in \mathfrak{p}$ is the unique vector for which $\partial_1\tau(X) = (\operatorname{Ad}_g X)_{\mathcal{M}}(g\mathcal{K})$ $= v_{g\mathcal{K}}$. We shall provide a proof of this statement later on directly from the Euler–Arnold equation, using the symmetries on the inner product.

3.1.3.1.1 Geodesics of $\mathbf{SPD}(n)$

As an example, let us compute the geodesics of the space of positive definite matrices, $\tau : \mathrm{GL}(n) \to \mathbf{SPD}(n)$, which is naturally reductive and $\tau(g) = gg^T$. We know its geodesics are given by $\tau(ge^{tX}) = ge^{tX}e^{tX^T}g^T = ge^{2tX}g^T$ for $(g,X) \in \mathrm{GL}(n) \times \mathfrak{p}$, where $\mathfrak{p}$ is the space of symmetric matrices. This has initial position gg^T and velocity $2gXg^T$. Writing $\Sigma \equiv gg^T$ implies $g = \Sigma^{\frac{1}{2}}$, and setting $V = 2gXg^T$ implies $2X = g^{-1}Vg^{-T}$. By choosing a representative $g \in g\mathcal{K}$ s.t., g is symmetric, we recover the formula used in [162] for the geodesics starting at $(\Sigma, V) \in T\mathbf{SPD}(n)$: $\tau(ge^{tX}) = \Sigma^{\frac{1}{2}}e^{t\Sigma^{-1/2}V\Sigma^{-1/2}}\Sigma^{\frac{1}{2}}$. In our implementation of HMC on $\mathbf{SPD}(n)$ via symplectic reduction, by working directly on $\mathrm{GL}(n)$ we will instead only need to compute ge^{tX} for the geodesic update given (g,X) (and will not need to update the velocity).

3.1.3.2 Geodesic Orbit Manifolds

More generally, consider a reductive manifold with $\mathcal{G}$-invariant metric as in §3.1.2. A geodesic γ starting at $g\mathcal{K}$ is called homogeneous if there exists a **geodesic vector** $X \in \mathfrak{g}$ which satisfies $\gamma(t) = \tau(e^{tX}g) = \tau(ge^{t\operatorname{Ad}_{g^{-1}}X})$. Physically, these geodesics can be interpreted as relative

equilibriums of the mechanical system [10]. Homogeneous geodesics are characterised by the geodesic lemma, which states that $0 \neq X \in \mathfrak{g}$ is a geodesic vector iff $\langle [X, Y]_\mathfrak{p}, X_\mathfrak{p} \rangle = 0$ for all $Y \in \mathfrak{p}$ [199] (recall $X_\mathfrak{p}$ denotes the projection of X onto $\mathfrak{p}$). A **geodesic orbit (g.o.) manifold** is one for which all the geodesics are homogeneous. It follows that naturally reductive and (Riemannian) symmetric spaces are g.o. manifolds. Finding formulae for the geodesics on a g.o. manifold (that is finding their geodesic vectors) reduces to solving linear equations for the coefficients of a basis of $\mathfrak{p}$ and $\mathfrak{k}$ in the reductive decomposition, as explained in [20].

3.2 Measure-Preserving Flows with Mechanics

We now introduce the mechanical systems of interests in HMC. We will begin by introducing the geometric description of dynamical systems and invariant measures in §3.2.1, which will be particularly useful in §4, and Hamiltonian mechanics in §3.2.2. Usually we have a metric on $\mathcal{M}$ and it is convenient to work in velocity space using the musical isomorphisms as we discuss in §3.2.3. In statistics, the Hamiltonian mechanics in velocity space defined by the metric has been described as "Lagrangian" and non-symplectic [202, 160]. In §3.2.4 we will discuss the Lagrangian perspective in order to correct these statements.

3.2.1 Dynamical Systems

A **dynamical system** on a manifold $\mathcal{F}$ is a C^1 vector field X. Given an initial condition $z \in \mathcal{F}$, the dynamical trajectories are the integral curves X, and in terms of the flow Φ^X these satisfy

$$\frac{\mathrm{d}\Phi_t^X(z)}{\mathrm{d}t} = X\left(\Phi_t^X(z)\right).$$

A volume form $\mu_\mathcal{F}$ (or one density) is **invariant under the dynamics** if for all t

$$(\Phi_t^X)^* \mu_\mathcal{F} = \mu_\mathcal{F},$$

which is equivalent to the flow preserving the associated measure $\mu_\mathcal{F}$ [3, proposition 8.1.11]:

$$(\Phi_t^X)_\sharp \mu_\mathcal{F} = \mu_\mathcal{F}, \qquad \forall t.$$

In general the relation between $(\Phi_t^X)^* \mu_\mathcal{F}$ and $\mu_\mathcal{F}$ is defined by the **Jacobian** $\mathcal{J}(\mu_\mathcal{F}, \Phi_t^X) : \mathcal{F} \to \mathbb{R}$

$$(\Phi_t^X)^* \mu_\mathcal{F} \equiv \mathcal{J}(\mu_\mathcal{F}, \Phi_t^X) \mu_\mathcal{F},$$

from which it follows that $\mathcal{J}(\mu_{\mathcal{F}}, \Phi_t^X)(z) = \det \partial_z \Phi_t^X$, where the determinant is defined with respect to $\mu_{\mathcal{F}}$. Another measure of the invariance of the dynamics is provided by the **compressibility** $\kappa : \mathcal{F} \to \mathbb{R}$ of the flow with respect to $\mu_{\mathcal{F}}$, defined as the divergence

$$\kappa \equiv \operatorname{div}(X), \qquad \text{where} \qquad \mathcal{L}_X \mu_{\mathcal{F}} \equiv \operatorname{div}(X) \mu_{\mathcal{F}}.$$

The Jacobian and compressibility are related by

$$\frac{\mathrm{d}}{\mathrm{d}t} \log \mathcal{J}(\mu_{\mathcal{F}}, \Phi_t^X) = \kappa \circ \Phi_t^X, \qquad \text{so formally} \qquad \mathcal{J}(\mu_{\mathcal{F}}, \Phi_t^X)(z) = e^{\int_0^t \kappa(\Phi_s^X(z))\mathrm{d}s},$$

from which it follow that the flow is **incompressible**, i.e., $\kappa = 0$, iff $\mathcal{J}(\mu_{\mathcal{F}}, \Phi_t^X) = 1$ iff $\mu_{\mathcal{F}}$ is preserved by the dynamics.

Non-Hamiltonian dynamical systems arise in statistical physics [254, 232, 343, 109], for example Nosé–Hoover dynamics, or systems with friction. These non-Hamiltonian dynamics have also been employed in the context of Monte Carlo methods [114]. Various formalisms (variational principles, brackets, transport theorems...) have been proposed to derive and analyse non-Hamiltonian flows as well as constructing invariant measures, see [346, 313, 344, 113, 314, 336, 61]. The construction of invariant measures often relies on the observation that if $\nu \equiv h\mu_{\mathcal{F}}$, then the Jacobians are related by

$$(\Phi_t^X)^* \nu = \mathcal{J}(\nu, \Phi_t^X)\nu = \frac{h \circ \Phi_t^X}{h} \mathcal{J}(\mu_{\mathcal{F}}, \Phi_t^X)\nu$$

so that if we choose h s.t.,

$$h\left(\Phi_t^X(z)\right) = h(z) e^{-\int_0^t \kappa(\Phi_s^X(z))\mathrm{d}s}$$

then ν is preserved by the dynamics. Following [346], if there exists a function f s.t., $\mathrm{d}f(X) = \kappa$, then $\int_0^t \kappa(\Phi_s^X(z))\mathrm{d}s = \int_0^t \mathrm{d}_{\Phi_s^X(z)} f(X_{\Phi_s^X(z)})\mathrm{d}s = \int_0^t (\Phi^X(z))^* \mathrm{d}f = \int_z^{\Phi_t^X(z)} \mathrm{d}f = f(\Phi_t^X(z)) - f(z)$ and $h \equiv e^{-f}$ solves the previous equation. It follows that

$$\nu \equiv e^{-f} \mu_{\mathcal{F}}, \qquad \text{where} \qquad \mathrm{d}f(X) = \kappa$$

is invariant under the flow of X. In fact this follows immediately from the proof of lemma 12, which shows $\operatorname{div}_{h\mu_{\mathcal{F}}}(X) = \operatorname{div}_{\mu_{\mathcal{F}}}(X) - \mathrm{d} \log |h|(X)$. In §4 we will consider non-Hamiltonian mechanical systems induced by a bracket, and characterise under what conditions they retain some of the key properties of Hamiltonian systems, and in particular show that these preserve a given measure iff the bracket is locally curled.

3.2.2 Hamiltonian Mechanics

Given a connected symplectic manifold $(\mathcal{F}, \omega)$ and a Hamiltonian $H : \mathcal{F} \to \mathbb{R}$, **Hamilton's equation** for $\hat{H} \in \Gamma(T\mathcal{F})$ is

$$i_{\hat{H}}\omega = \mathrm{d}H. \tag{3.1}$$

In local Darboux coordinates $z = (q, p)$,

$$\omega = \mathrm{d}q^i \wedge \mathrm{d}p_i, \qquad \hat{H} = \frac{\partial H}{\partial p_i}\frac{\partial}{\partial q^i} - \frac{\partial H}{\partial q^i}\frac{\partial}{\partial p_i}$$

and if $t \mapsto z(t)$ is an integral curve of $\hat{H}$ [3]

$$\dot{z} = \begin{pmatrix} \dot{q} \\ \dot{p} \end{pmatrix} = \begin{pmatrix} \nabla_p H \\ -\nabla_q H \end{pmatrix} = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix} \nabla_z H \equiv J\nabla_z H. \tag{3.2}$$

Hamilton's equation (3.1) defines a bijection between Hamiltonian vector fields and exact 1-forms, and thus a 1-1 correspondence between Hamiltonian vector fields and functions up to a constant: $\hat{H} = \widehat{H + c}$ for any $c \in \mathbb{R}$. This invariance of Hamiltonian vector fields with respect to the transformation $H \mapsto H + c$ is fundamental for applications of mechanics to sampling as it implies we only need to know the Hamiltonian up to constant, and thus the target density up to normalisation. We will sometimes use the **Poisson brackets** $\{\cdot, \cdot\} : \mathcal{F} \times \mathcal{F} \to \mathcal{F}$, defined by $\{f, h\} \equiv -\omega(\hat{f}, \hat{g})$. These have the property that $\{f, h\} = 0$ iff Φ_t^f and Φ_t^h commute, so the Poisson bracket measures the non-commutativity of the Hamiltonian flows (5.3 [212]). If ω is exact, i.e., there exists a 1-form Θ s.t., $\omega = -\mathrm{d}\Theta$, we call Θ the **Poincaré/Liouville 1-form**. This is the case for the cotangent bundle $T^*\mathcal{M}$ of a manifold $\mathcal{M}$. In that case the cotangent-lifted coordinates (q^i, p_j) are Darboux, and the Poincaré 1-form is locally given by $\Theta = p_i \mathrm{d}q^i$ [164]. Moreover observe

$$X_H(H) = \omega(X_H, X_H) = 0, \quad \mathcal{L}_{\hat{H}}\omega = i_{\hat{H}}\mathrm{d}\omega + \mathrm{d}i_{\hat{H}}\omega = 0 = \mathcal{L}_{\hat{H}}\omega^n, \tag{3.3}$$

that is the Hamiltonian flow preserves the energy H, the symplectic structure, and the symplectic volume ω^n. Hence, from §3.2.1, the Hamiltonian flow Φ^H of $\hat{H}$ is incompressible with respect to ω^n, and preserves the **canonical distribution** $\mu_H \equiv e^{-H}\omega^n / \int e^{-H}\omega^n$

$$(\Phi_t^H)_\sharp \mu_H(A) \equiv \int_{(\Phi_t^H)^{-1}(A)} \mu_H = \int_A (\Phi_{-t}^H)^* (\mu_H) = \int_A \mu_H = \mu_H(A),$$

[3] Note we also have $\omega(u, v) = u^\top J v$

which motivates the use of Hamiltonian flows as proposals to generate μ_H-preserving Markov chains. A diffeomorphism $f : (\mathcal{F}_1, \omega_1) \to (\mathcal{F}_2, \omega_2)$ will be said to be a **symplectomorphism** if $f^*\omega_2 = \omega_1$. Symplectomorphisms can be used to transport samples [255] between the canonical distributions μ_{f^*H} and μ_H, a powerful fact that allows the use of symplectic transformations to simplify the sampler by mapping it into coordinates in which the implementation is less expensive. For the purpose of sampling from a target P on a manifold $\mathcal{M}$, the importance of this result stems from the fact that, as we will explain in §3.3, we can express expectations with respect to P in terms of expectations with respect to an appropriately chosen canonical distribution. From above we know that by preservation of energy the flow remains stuck in a given energy level. As we now show in §3.2.2.1, there exists a microcanonical distribution on each energy levels set which is invariant under the restricted flow, and expectations with respect to μ_H may be themselves expressed as a weighted integral of microcanonical expectations.

3.2.2.1 Microcanonical Distribution

The microcanonical distribution plays a central role in HMC for molecular dynamics, and we now provide a geometric construction. Consider a Hamiltonian H on $(\mathcal{F}, \omega)$ and suppose that H is a submersion, so the energy level sets $H^{-1}(E)$ are submanifolds for any $E \in \mathbb{R}$. We define the **microcanonical volume form** δ_E on a connected component $\Sigma_E \subset H^{-1}(E)$ to be the restriction of any form σ satisfying $\omega^n = \mathrm{d}H \wedge \sigma$, so $\delta_E \equiv \iota_E^*\sigma$ where $\iota_E : \Sigma_E \hookrightarrow \mathcal{F}$ (this definition is independent of the choice of σ). In the following we assume $\Sigma_E = H^{-1}(E)$. By (3.3) the Hamiltonian H is preserved along the flow of X_H, so X_H restricts to a vector field $X_H|_{H^{-1}(E)}$ on $H^{-1}(E)$, which preserves the microcanonical volume form [2, 3.4]

$$\mathcal{L}_{X_H|_{H^{-1}(E)}}\delta_E = 0, \qquad \text{equivalently} \qquad (\Phi_t^H|_{H^{-1}(E)})^*\delta_E = \delta_E \quad \forall t.$$

The fibre integral (see [59]) $\int_{fib}\sigma$ of σ is known as the **density of states** in statistical physics [185]. By definition its value at E, sometimes denoted $\mathrm{d}(E) \in \mathbb{R}$, is[4]

$$\left(\int_{fib}\sigma\right)(E) = \int_{H^{-1}(E)}\iota_E^*\sigma = \int_{H^{-1}(E)}\delta_E \equiv \mathrm{d}(E)$$

and it corresponds to the density of the pushforward $H_\sharp\omega^n$ with respect to the Lebesgue measure $\mathrm{d}E$ on $\mathbb{R}$, since

$$i_{\partial_E}H_\sharp\omega^n = i_{\partial_E}\int_{fib}\omega^n = \int_{fib}i_v\omega^n = \int_{fib}\mathrm{d}H(v)\sigma = \int_{fib}\sigma$$

[4] This definition is consistent with the one discussed in 2.1.1 [46], since if $X \in \Gamma(T\mathcal{F})$ satisfies $\mathrm{d}H(X) > 0$ $(\int_{H^{-1}(E)}\iota_E^*(i_X\omega^n))/\mathrm{d}H(X) = (\int_{H^{-1}(E)}\mathrm{d}H(X)\delta_E)/\mathrm{d}H(X) = \int_{H^{-1}(E)}\delta_E = \mathrm{d}(E)$.

where v is any lift of ∂_E, and we used that $\iota_E^* \mathrm{d}H = 0$ for any E (which implies $\int_{fib} \mathrm{d}H \wedge i_v \sigma = 0$), and $\mathrm{d}H(v) = \mathrm{d}E(\partial H(v)) = \mathrm{d}E(\partial_E) = 1$. The **microcanonical distribution** $\mu_{mc,E}$ is the associated probability measure on $H^{-1}(E)$

$$\mu_{mc,E} \equiv \delta_E / \int_{H^{-1}(E)} \delta_E.$$

An important result is that the map $E \mapsto \delta_E$ disintegrates the symplectic volume form with respect to H and $\mathrm{d}E$, and thus we can write canonical expectations in terms of microcanonical expectations: using $\mathrm{d}H = H^* \mathrm{d}E$ and the properties of the fibre integral [3],

$$\begin{aligned}\int_{\mathcal{F}} f e^{-H} \omega^n &= \int_{\mathcal{F}} f e^{-H} \mathrm{d}H \wedge \sigma = \int_{\mathcal{F}} f e^{-H} H^* \mathrm{d}E \wedge \sigma = \int_{\mathbb{R}} \mathrm{d}E \int_{H^{-1}(E)} f e^{-H} \delta_E \\ &= \int_{\mathbb{R}} e^{-E} \mathrm{d}(E) \mathrm{d}E \int_{H^{-1}(E)} f \mu_{mc,E},\end{aligned}$$

and thus

$$\mathbb{E}_{\mu_H}[f] = \frac{1}{\int_{\mathcal{F}} e^{-H} \omega^n} \int_{\mathbb{R}} \mathbb{E}_{\mu_{mc,E}}[f] e^{-E} \mathrm{d}(E) \mathrm{d}E.$$

The idea of expressing the expectation with respect to an unconstrained measure in terms of the expectation of a constrained measure (as above) appears for example in Blue Moon samplers for molecular systems [73]. As mentioned above the microcanonical volume form is a disintegration. Disintegrations are usually called "delta measures" in statistical physics, and are often expanded in terms of "sigma measures" via the coarea formula. Geometrically, sigma measures correspond to a choice of smooth measure on the fibre, which is necessary to define the co-Jacobian of the coarea formula [84]. For example when $\mathcal{F} = T^*\mathbb{R}^n \cong \mathbb{R}^{2n}$, we can express the microcanonical distribution in terms of the Riemannian surface measure $\mathrm{vol}_{H^{-1}(E)}$ on $H^{-1}(E)$ induced by the ambient Euclidean inner product (which is a natural choice of sigma measure in this case). Then by the coarea formula

$$\mu_{mc,E} = \frac{1}{\mathrm{d}(E)} \frac{1}{\|\nabla H\|} \mathrm{vol}_{H^{-1}(E)}. \tag{3.4}$$

3.2.2.2 Invariant Measures on Invariant Manifolds

We can generalise the setting of §3.2.2.1 to the case where a Hamiltonian vector field X_H preserves m functions $f_i : \mathcal{F} \to \mathbb{R}$, i.e., $X_H(f_i) = 0$ (or equivalently $\{H, f_i\} = 0$), representing m conserved quantities. Assume $f \equiv (f_1, \ldots, f_m) : \mathcal{F} \to \mathbb{R}^m$ is a submersion, so that $f^{-1}(E)$ is a submanifold. By theorem 3.4.15 [2] X_H preserves $f^{-1}(E)$, i.e., $\Phi_t^H f^{-1}(E) \subset f^{-1}(E)$, and in a neighbourhood $f^{-1}(E)$ we have $\omega^n = \mathrm{d}f_1 \wedge \cdots \wedge \mathrm{d}f_m \wedge \sigma$ for some differential form σ.

Then the measure $\iota_E^*\sigma$ on $f^{-1}(E)$ is preserved by the flow $\Phi_t^H|_{f^{-1}(E)}$ and is independent of the choice of σ. Using the results form §3.3, denoting by $E_i : \mathbb{R}^m \to \mathbb{R}$ the coordinates, and setting $\mu_{mc,E} \equiv \iota_E^*\sigma / \int_{fib} \sigma(E)$

$$\begin{aligned}\mathbb{E}_{\mu_H}[h] &= \int_{\mathbb{R}^m} \circ \int_{fib} he^{-H}\omega^n = \int_{\mathbb{R}^m} \circ \int_{fib} he^{-H} f^* \left(\mathrm{d}E_1 \wedge \cdots \wedge \mathrm{d}E_m\right) \wedge \sigma \\ &= \int_{\mathbb{R}^m} \mathrm{d}E_1 \wedge \cdots \wedge \mathrm{d}E_m \circ \int_{fib} he^{-H}\sigma = \int_{\mathbb{R}^m} \frac{\mathrm{d}E}{\int_{fib}\sigma(E)} \mathbb{E}_{\mu_{mc,E}}[he^{-H}]\end{aligned}$$

3.2.3 Hamiltonian Mechanics on $T\mathcal{M}$: Riemannian Newton Equations

While in physics mechanics is described in momentum space, when implementing HMC with geodesic flows, the velocity space version of mechanics (on $T\mathcal{M}$) may be convenient, and we shall now derive it. Let $\langle \cdot, \cdot \rangle$ be a Riemannian metric on $\mathcal{M}$ and consider the musical isomorphism $\flat : T\mathcal{M} \to T^*\mathcal{M}$, $v \mapsto v^\flat \equiv \langle v, \cdot \rangle$. On $T\mathcal{M}$ we define the symplectic structure $\omega_\flat \equiv \flat^*\omega$. Consequently $T\mathcal{M}$ and $T^*\mathcal{M}$ are symplectomorphic, and

$$\begin{aligned} i_{\hat{H}}\omega = \mathrm{d}H \quad &\Leftrightarrow \quad \flat^* i_{\hat{H}}\omega = \flat^*\mathrm{d}H \quad \Leftrightarrow \quad \omega(\hat{H}, \flat_* \cdot) = \mathrm{d}\flat^* H \\ &\Leftrightarrow \quad \omega_\flat((\flat^{-1})_*\hat{H}, \cdot) = \mathrm{d}\flat^* H \quad \Leftrightarrow \quad i_{(\flat^{-1})_*\hat{H}}\omega_\flat = \mathrm{d}\flat^* H. \end{aligned}$$

It follows that $X_H \in \Gamma(TT\mathcal{M})$ is the Hamiltonian vector field of $H : T\mathcal{M} \to \mathbb{R}$ iff $\flat_* X_H$ is the Hamiltonian vector field of $H \circ \flat^{-1}$: $\flat_* X_H = X_{H \circ \flat^{-1}}$. Similarly, the canonical distributions satisfy $\flat_\sharp \left(e^{-H}\omega_\flat^n\right) = e^{-H\circ\flat^{-1}}\omega^n$. Note however that, while in cotangent-lifted coordinates the symplectic structure takes the canonical form $\omega = \mathrm{d}q^i \wedge \mathrm{d}p_i$, in tangent-lifted coordinates, since $p_i = G_{ik}v^k$ where $G_{ik} \equiv \left\langle \partial_{q^i}, \partial_{q^k} \right\rangle$, we have

$$\omega_\flat = v^k \partial_{q^j} G_{ik} \mathrm{d}q^i \wedge \mathrm{d}q^j + G_{ij}\mathrm{d}q^i \wedge \mathrm{d}v^j, \tag{3.5}$$

and this seems to be the origin of some confusion in HMC. The tangent-lifted coordinates (q^i, v^i) are not a canonical choice of coordinates for the symplectic structure, but by Darboux's theorem there exists local coordinates in which $\omega_\flat$ takes the usual form $\mathrm{d}\tilde{q}^i \wedge \mathrm{d}\tilde{v}^i$. Treating the symplectic manifold $(T\mathcal{M}, \omega_\flat)$ as a Hamiltonian system in its own right, we can as usual derive Hamilton's equations in local coordinates from $\omega_\flat(X, \cdot) = \mathrm{d}H(\cdot)$. The local Hamilton's equations are

$$X^i \equiv \mathrm{d}q^i(X) = G^{is}\partial_{v^s} H, \tag{3.6a}$$

$$\overline{X}^r \equiv \mathrm{d}v^r(X) = G^{ir}v^k X^j \left(\partial_{q^i} G_{jk} - \partial_{q^j} G_{ik}\right) - G^{rs}\partial_{q^s} H. \tag{3.6b}$$

Proof. Recall that if (q, v) are tangent-lifted coordinates on $T\mathcal{M}$, then $(\partial_{v^i}|_q)$ forms a basis of the vertical space $V_q T\mathcal{M}$, and $\left(\partial_{q^i}|_q, \partial_{v^i}|_q\right)$ for a basis of $T_q T\mathcal{M}$. Let $Y = Y^i \partial_{q^i} + \overline{Y}^i \partial_{v^i} \in \Gamma(TT\mathcal{M})$. We want to find the vector field $X = X^i \partial_{q^i} + \overline{X}^i \partial_{v^i}$ that solves Hamilton's equations. Then

$$\mathrm{d}H(Y) = Y(H) = Y^i \partial_{q^i} H + \overline{Y}^i \partial_{v^i} H$$

For the LHS we recall $\mathrm{d}q^i \wedge \mathrm{d}q^j \equiv \mathrm{d}q^i \otimes \mathrm{d}q^j - \mathrm{d}q^j \otimes \mathrm{d}q^i$, so

$$\omega_\flat(X, Y) = v^k \partial_{q^j} G_{ik} \left(X^i Y^j - X^j Y^i\right) + G_{ij} \left(X^i \overline{Y}^j - \overline{X}^j Y^i\right)$$

equating the coefficients of $\overline{Y}^s$ implies $X^i = G^{is} \partial_{v^s} H$, and similarly for the coefficients of Y^i:

$$\overline{X}^r = G^{ir} v^k X^j \left(\partial_{q^i} G_{jk} - \partial_{q^j} G_{ik}\right) - G^{rs} \partial_{q^s} H.$$

■

In particular, if $H(q, v) = \frac{1}{2} G_{ij} v^i v^j + V(q)$, we recover the standard system

$$X^i = v^i, \qquad \overline{X}^r = -\Gamma^r_{ik} v^i v^k - G^{rs} \partial_{q^s} V \tag{3.7}$$

where Γ^r_{ik} are the components of connection compatible with the Riemannian metric (the Levi-Civita connection), though this form is not as useful for HMC as X_H is usually not tractable;

Proof. If $H(q, v) = \frac{1}{2} G_{ij} v^i v^j + V(q)$, then $\partial_{v^s} H = G_{is} v^i$, $\partial_{q^s} H = \frac{1}{2} \partial_{q^s} G_{ij} v^i v^j + \partial_{q^s} V$. Hence $X^i = v^i$, and

$$\begin{aligned}
\overline{X}^r &= G^{ir} v^k X^j \left(\partial_{q^i} G_{jk} - \partial_{q^j} G_{ik}\right) - \tfrac{1}{2} G^{rs} \partial_{q^s} G_{ij} v^i v^j - G^{rs} \partial_{q^s} V \\
&= G^{ir} v^k v^j \left(\partial_{q^i} G_{jk} - \partial_{q^j} G_{ik} - \tfrac{1}{2} \partial_{q^i} G_{kj}\right) - G^{rs} \partial_{q^s} V = G^{ir} v^k v^j \left(\tfrac{1}{2} \partial_{q^i} G_{jk} - \partial_{q^j} G_{ik}\right) - G^{rs} \partial_{q^s} V \\
&= G^{ir} v^k v^j \left(\tfrac{1}{2} \partial_{q^i} G_{jk} - \tfrac{1}{2} \partial_{q^j} G_{ik} - \tfrac{1}{2} \partial_{q^k} G_{ij}\right) = -v^k v^j \Gamma^r_{kj} - G^{rs} \partial_{q^s} V.
\end{aligned}$$

■

These are often expressed in terms of an integral curve $\gamma(t) = (q^i(t), v^k(t))$ of X

$$\dot{q} = v, \qquad \dot{v}^k = -\Gamma^k_{ij} v^i v^j - G^{ki} \partial_{q^i} V,$$

or in a coordinate-independent manner as

$$\frac{\nabla \dot{\gamma}}{\mathrm{d}t} = -\nabla V, \quad \text{or} \quad \flat \frac{\nabla \dot{\gamma}}{\mathrm{d}t} = -\mathrm{d}V,$$

where ∇ is the Riemannian gradient. Comparing the latter with (2.2), we see we have constructed a curved analogue of Newton's equations. Note that in tangent-lifted coordinates, the symplectic measure is

$$\omega_\flat^n = \det(G) \mathrm{d}q \wedge \mathrm{d}v.$$

3.2.4 Lagrangian Mechanics

In the previous sections we have introduced the Hamiltonian description of mechanics, which relies on symplectic geometry. In statistics, Hamiltonian mechanics in velocity space is sometimes referred to as (geodesic) Lagrangian Monte Carlo (LMC) [202, 204, 161]. It is sometimes claimed that LMC "cannot be recognised as a Hamiltonian system" and "non-symplectic" (and thus requires Jacobians corrections) [202, 160, 161], but as shown above $T\mathcal{M}$ really is a symplectic manifold, and in particular the Hamiltonian flows are automatically volume-preserving. In fact observe that no Lagrangian is involved in the construction of $\omega_\flat$, the latter relies only on the metric. In §A.5.3 we recall the Lagrangian formulation which is based on the principle of stationary action (and so is the Lagrangian theory of numerical integrators). If we do not have a metric, but are given a (regular) Lagrangian $\mathcal{L}$, we can still induce a symplectic geometry on $T\mathcal{M}$ under which Lagrangian flows are symplectic since they may be obtained as a solution of Hamilton's equations, as we now describe in §3.2.4.1.

3.2.4.1 Lagrangian Symplectic Structure

Given a **regular** Lagrangian $\mathcal{L}$, i.e., one for which the fibre derivative $\mathbb{F}\mathcal{L} : T\mathcal{M} \to T^*\mathcal{M}$ is locally invertible, it is also possible to define the **Lagrangian two-form** $\omega_\mathcal{L} \equiv \mathbb{F}\mathcal{L}^*\omega$ on $T\mathcal{M}$, which in local coordinates is

$$\omega_\mathcal{L} = \frac{\partial^2 \mathcal{L}}{\partial q^i \partial v^j} \mathrm{d}q^i \wedge \mathrm{d}q^j + \frac{\partial^2 \mathcal{L}}{\partial v^i \partial v^j} \mathrm{d}q^j \wedge \mathrm{d}v^i. \tag{3.8}$$

Defining the **Lagrangian energy** $E(v) \equiv \langle \mathbb{F}\mathcal{L}(v), v \rangle_* - \mathcal{L}(v)$ [249], we call Z a **Lagrangian vector field** if it is a solution to Hamilton's equation $\omega_\mathcal{L}(Z, \cdot) = \mathrm{d}E$. This is equivalent to demanding that the integral curves of Z solve Euler-Lagrange equations §A.5.3. Hence Lagrangian flows are symplectic, a fact first discovered by Lagrange himself, and preserve the "canonical distribution" $\propto e^{-E}\omega_\mathcal{L}^n$. Note that if $\mathcal{L}$ is hyperregular (i.e., $\mathbb{F}\mathcal{L}$ is a diffeomorphism), then by definition a Lagrangian vector field is $\mathbb{F}\mathcal{L}$-related to the Hamiltonian vector field of $H(p) \equiv E \circ \mathbb{F}\mathcal{L}^{-1}(p) = \langle p, \mathbb{F}\mathcal{L}^{-1}(p) \rangle_* - \mathcal{L}(\mathbb{F}\mathcal{L}^{-1}(p))$ on $T^*\mathcal{M}$. The map $\mathcal{L} \mapsto H \equiv E \circ \mathbb{F}\mathcal{L}^{-1}$ is called the **Legendre transformation** of $\mathcal{L}$ (although it can also refer to the map $v \mapsto \mathbb{F}\mathcal{L}(v)$ in the literature).

As mentioned above the construction of the symplectic manifold $(T\mathcal{M}, \omega_{\mathcal{L}})$ is based on a fixed Lagrangian, and does not depend on whether $\mathcal{M}$ is given a Riemannian metric. If we have both a regular Lagrangian and Riemannian metric on $\mathcal{M}$, then $\omega_{\mathcal{L}} = (\flat^{-1} \circ \mathbb{F}\mathcal{L})^* \omega_\flat$. For a standard Lagrangian $\mathcal{L} \equiv \frac{1}{2}\|\cdot\|^2 - V \circ \pi_{T\mathcal{M}} : T\mathcal{M} \to \mathbb{R}$, where $\|\cdot\|$ is the norm associated to the metric, then $\mathbb{F}\mathcal{L}(v) = \langle v, \cdot \rangle = \flat(v)$, so $\omega_\flat = \omega_{\mathcal{L}}$, and the Legendre transform of $\mathcal{L}$ is $H \equiv \frac{1}{2}\|\cdot\|^2 + V \circ \pi_{T^*\mathcal{M}}$, where now $\|\cdot\|$ is the norm of the dual Riemannian metric. However, as we will discuss later, to construct numerical integrators for HMC we often split the Hamiltonian $H = V \circ \pi + \frac{1}{2}\|\cdot\|^2$ into the Hamiltonians $H_1 \equiv V \circ \pi$ and $H_2 \equiv \frac{1}{2}\|\cdot\|^2$ in order to approximate the flow of $\hat{H}$, and even though H, H_2 are hyperregular, H_1 is not, so does not correspond to a Lagrangian system (for H_1 we have $\mathbb{F}H_1(p) = 0$). Indeed splitting methods are specific to the Hamiltonian description: the Lagrangian perspective instead relies on the discrete variational principle.

3.3 Construction of Hamiltonian System Adapted to Target Distribution

Now that we have introduced mechanics and its measure-preserving properties, we need to connect these mechanical systems to our statistical problem. In order to sample from a target distribution P on $\mathcal{M}$, HMC algorithms rely on constructing a mechanical system on a fibre bundle $\pi : \mathcal{F} \to \mathcal{M}$ over $\mathcal{M}$, where $(\mathcal{F}, \omega)$ is a symplectic manifold [221], and a canonical distribution μ_H satisfying $\pi_\sharp \mu_H = P$. Typically $(\mathcal{F}, \omega)$ will be an **exact symplectic fibre bundle** over $\mathcal{M}$, so $\omega = -\mathrm{d}\Theta$ where $\Theta \in \Gamma(\mathcal{F}, T^*\mathcal{F})$ is a Liouville form as defined in 1 [11], and 2 [220]. The most important cases are $\mathcal{F} = T^*\mathcal{M}$, $\mathcal{F} = \mathcal{G} \times \mathfrak{g}^*$, $\mathcal{F} = T\mathcal{M}$ when $\mathcal{M}$ is Riemannian or has a regular Lagrangian, and $\mathcal{F} = \mathcal{G} \times \mathfrak{g}$ when $\mathfrak{g}$ has an inner product (here $\mathcal{G}$ is a Lie group with Lie algebra $\mathfrak{g}$). Local expressions for their symplectic structures are provided in §3.3.1. Then let $\zeta \in \Omega^n(\mathcal{F})$ be any n-form s.t., $\iota_q^* \zeta$ is a volume form on $\pi^{-1}(q)$ (where $\iota_q : \pi^{-1}(q) \hookrightarrow \mathcal{F}$ is the inclusion), and whose fibre integral is one, $\int_{fib} \zeta = 1$, i.e.,

$$\int_{fib} \zeta(q) = \int_{\pi^{-1}(q)} \iota_q^* \zeta = 1, \qquad \forall q \in \mathcal{M}.$$

Then ζ is a smooth **probability disintegration** (or regular conditional distribution) of the smooth probability measure $\pi^* P \wedge \zeta$ on $\mathcal{F}$ with respect to π and P, in particular $\pi_\sharp(\pi^* P \wedge \zeta) =$

P, as we will see below using (see 8 [3])

$$\int_{fib} \pi^* P \wedge \zeta = P \wedge \int_{fib} \zeta, \quad \text{and} \quad \int_{\mathcal{F}} = \int_{\mathcal{M}} \circ \int_{fib}.$$

Since we can build Hamiltonian flows on $\mathcal{F}$ that preserve the canonical distribution μ_H (recall §3.2.2), it is convenient to express $\pi^* P \wedge \zeta$ as a canonical distribution by defining

$$H \equiv -\log \frac{\mathrm{d}(\pi^* P \wedge \zeta)}{\mathrm{d}\omega^n},$$

since then

$$\pi^* P \wedge \zeta = e^{-H} \omega^n \equiv \mu_H.$$

By construction $\pi_\sharp \mu_H = P$. Indeed for any $f \in L^1(P)$

$$\begin{aligned} \pi_\sharp \mu_H[f] &= \int_{\mathcal{F}} \pi^* f \mu_H = \int_{\mathcal{F}} \pi^*(fP) \wedge \zeta = \int_{\mathcal{M}} \circ \int_{fib} (\pi^*(fP) \wedge \zeta) \\ &= \int_{\mathcal{M}} \left(fP \wedge \int_{fib} \zeta \right) = \int_{\mathcal{M}} fP \int_{fib} \zeta = \int fP = P[f]. \end{aligned}$$

Note this simplicity and clarity of the above derivation in terms of the fibre integral: in just two lines we have been able to obtain the main results of [44], and at the right level of generality since several HMC schemes, in particular the ones from lattice QCD, are not implemented on the cotangent bundle. Using [84] we can also work with one densities, by viewing $\pi^* P \wedge \zeta$ as a "retrenchment", and we will often use this fact in the next chapter. We have thus constructed a Hamiltonian system $(\mathcal{F}, \omega, H)$, called **adapted to** P, with the property that the pushforward of the canonical distribution μ_H with respect to π is precisely the target distribution P. Hence if $Z_i \sim \mu_H$ are independent μ_H-distributed random variables, then $Q_i \equiv \pi \circ Z_i$ are independent P-distributed random variables, so that if we can sample from μ_H, then the projected samples will be samples from P. In §3.4.1 we will discuss how to obtain samples from μ_H by combining the Hamiltonian flow Φ_t^H, which preserves μ_H but remains stuck in the level sets of H (see §3.2.2.1), with a thermostat process which varies the energy level sets. Let us now provide the general expression for the symplectic structure with respect to a local frame (which allows us to derive equations of motion in that frame). In §3.3.2 and §3.3.3 we will then construct Hamiltonian systems adapted to appropriate targets on Riemannian manifolds and Lie groups.

3.3.1 Symplectic Structures on Fibre Bundles

Following [11] we can obtain a general local expression for the symplectic structure of $(\mathcal{F}, -\mathrm{d}\Theta)$ in a local trivialisation $X_i \in \Gamma(T\mathcal{M})$ of $\mathcal{M}$, with α^i its dual frame, $\alpha^j(X_k) = \delta_k^j$. Since Θ is a

horizontal form, we can expand it as $\Theta = p_i \pi^* \alpha^j$ for some local functions $p_i \in C^\infty(\mathcal{F})$, and

$$\omega = -\mathrm{d}\Theta = -\mathrm{d}p_i \wedge \pi^*\alpha^i - p_i \wedge \pi^*\mathrm{d}\alpha^i = \pi^*\alpha^i \wedge \mathrm{d}p_i + \tfrac{1}{2} p_i \pi^* \left(c^i_{jk} \alpha^j \wedge \alpha^k \right)$$

where $c^r_{jk} \in C^\infty(\mathcal{M})$ are given by $[X_i, X_j] \equiv c^k_{ij} X_k$, and the last equality follows from $\mathrm{d}\beta(X,Y) = X[\beta(Y)] - Y[\beta(X)] - \beta\big([X,Y]\big)$.

(i) In the case $\mathcal{F} = T^*\mathcal{M}$, if q^i are local coordinates on $U \subset \mathcal{M}$, then we can set $X_i = \partial_{q^i}$, so $\alpha^i = \mathrm{d}q^i$, and $p_i : \pi^{-1}(U) \to \mathbb{R}$ the cotangent-lifted coordinates, $p_i(\alpha) = \alpha(\partial_{q^i})$. Then $\Theta = p_i \pi^* \mathrm{d}q^i$ is the usual Liouville form. Note $c^i_{jk} = 0$, so $\omega = \pi^*\mathrm{d}q^i \wedge \mathrm{d}p_i$.

(ii) If $\mathcal{M}$ is a Riemannian manifold, and q^i are local coordinates, $X_i = \partial_{q^i}$, the tangent-lifted coordinates are $v^i : X \mapsto \mathrm{d}q^i(X)$. We have $\Theta_\flat = G_{ij} v^j \pi^* \mathrm{d}q^i$ where $G_{ik} \equiv \langle \partial_{q^i}, \partial_{q^k} \rangle$ (see (7.2.9) [251]), and $c^i_{jk} = 0$, so we recover (3.5).

(iii) On a Lie group we can find global expressions for Θ and ω. On $\mathcal{F} = T^*\mathcal{G}$ the Liouville form can be written as $\Theta = J\,(\pi^*\theta)$, where $\theta \in \Gamma(\mathcal{G}, T^*\mathcal{G} \otimes \mathfrak{g})$ is the Maurer-Cartan form, and $J : T^*\mathcal{G} \to \mathfrak{g}^*$ the momentum map of the right action of $\mathcal{G}$ on itself, $J(\alpha_g) = L^*_g \alpha_g$ for any $\alpha_g \in T^*_g\mathcal{G}$ (see §3.9). If (θ^i) denotes a basis of $\mathfrak{g}^*_L$, we can expand J and θ in this frame (so $J = p_i \theta^i$, for $p_i : T^*\mathcal{G} \to \mathbb{R}$) to obtain, $\Theta = p_i \pi^* \theta^i$, and

$$\omega = \underbrace{\pi^*\theta^i \wedge \mathrm{d}p_i}_{\text{usual ``d}x \wedge \mathrm{d}p\text{'' term}} + \underbrace{\tfrac{1}{2} p_i c^i_{jk} \pi^*\theta^j \wedge \pi^*\theta^k}_{\text{additional non-abelian term}}$$

where c^i_{jk} are the structure constant of the Lie algebra determined by the dual basis to (θ^i). In particular on an abelian group such as $\mathcal{G} = \mathbb{R}^n$ we have $c^i_{jk} = 0$.

(iv) Given an inner product on $\mathfrak{g}$, we can consider the symplectic trivial bundle $\mathcal{F} = \mathcal{G} \times \mathfrak{g}$ over $\mathcal{G}$. Choose a basis (T_i) of $\mathfrak{g}$ and define the coordinates $v^i : \mathfrak{g} \to \mathbb{R}$, $v^i(X) \equiv \langle X, T_i \rangle_\mathfrak{g}$, and components of the inner product $G_{ij} \equiv \langle T_i, T_j \rangle_\mathfrak{g}$. Then (see [26])

$$\omega = G_{ij} \pi^*\theta^i \wedge \mathrm{d}v^i + \tfrac{1}{2} G_{ir} v^r c^i_{jk} \pi^*\theta^j \wedge \pi^*\theta^k.$$

(v) Let $\mathcal{L} : T\mathcal{M} \to \mathbb{R}$ be a regular Lagrangian (i.e., its fibre derivative $\mathbb{F}\mathcal{L}$ is locally invertible) on a manifold $\mathcal{M}$. The **Lagrangian 1-form** is defined by choosing $\alpha^i \equiv \mathrm{d}q^i$, and $p_i \equiv \frac{\partial \mathcal{L}}{\partial v^i}$, $\Theta_\mathcal{L} \equiv \frac{\partial \mathcal{L}}{\partial v^i} \pi^* \mathrm{d}q^i$. Then $\omega_\mathcal{L} \equiv -\mathrm{d}\Theta_\mathcal{L}$ is given by (3.8).

3.3.2 Riemannian Newton System

Consider the case where the target distribution is expressed using the Riemannian measure $\mathrm{vol}_{\mathcal{M}}$ on a Riemannian manifold $\mathcal{M}$, $P = \frac{1}{Z_P} e^{-V} \mathrm{vol}_{\mathcal{M}}$. In that case it is convenient to employ the Riemannian symplectic manifold $(\mathcal{F}, \omega) \equiv (T\mathcal{M}, \omega_\flat)$ discussed in §3.2.3. If we define the disintegration by setting $\zeta|_{T_q\mathcal{M}} \equiv \mathcal{N}(0, \langle \cdot, \cdot \rangle_q^{-1})$, then the canonical measure is chosen to correspond to the standard Hamiltonian $H(q, v) = V(q) + \frac{1}{2} \langle v, v \rangle_q$, since in local tangent-lifted coordinates, $\zeta(q, v) = \frac{1}{\sqrt{(2\pi)^n \det G(q)^{-1}}} \exp\left(-\frac{1}{2} v^\top G(q) v\right) \mathrm{d}v$[5], and

$$\begin{aligned}
\mu_H(q, v) \equiv \frac{1}{Z_P} \left(e^{-H} \omega_\flat^n\right)(q, v) &= \frac{1}{Z_P} e^{-V(q)} e^{-T(q,v)} \det G(q) \mathrm{d}q \wedge \mathrm{d}v \\
&= \frac{1}{Z_P} e^{-V(q)} \sqrt{\det G(q)} \mathrm{d}q \wedge e^{-T(q,v)} \sqrt{\det G(q)} \mathrm{d}v \\
&= \left(\frac{1}{Z_P} e^{-V} \mathrm{vol}_{\mathcal{M}}\right)(q) \wedge e^{-\frac{1}{2}\langle v, v \rangle_q} \sqrt{\det G(q)} \mathrm{d}v \\
&= (\pi^* P \wedge \zeta)(q, v),
\end{aligned}$$

so $\mu_H = \pi^* P \wedge \zeta$ as required. The Hamiltonian system $(T\mathcal{M}, \omega_\flat, H \equiv \pi^* V + \frac{1}{2} \| \cdot \|^2)$ will be called the **Riemannian Newton system adapted to** P.

3.3.3 Lie Group Hamiltonian System

Let $\mathcal{M} = \mathcal{G}$ be a Lie group with dual Lie algebra $\mathfrak{g}^*$. Typically $P = \frac{1}{Z_P} e^{-V} \theta^1 \wedge \cdots \wedge \theta^n$ where $\theta^1 \wedge \cdots \wedge \theta^n$ is the left Haar measure, and (θ^i) a basis of left-invariant 1-forms (see §3.1.1). In that case it is convenient to adopt the symplectic manifold $\mathcal{G} \times \mathfrak{g}^*$ mentioned in §3.3.1. Using the coordinates $p_i : \mathfrak{g}^* \to \mathbb{R}$ defined by $p_i(\alpha) = \alpha(T_i)$, where (T_i) is a basis of $\mathfrak{g}$ dual to $(\theta^i(1))$, we can identify $\mathfrak{g}^* \cong \mathbb{R}^n$. We define ζ to be a probability measure on $\mathbb{R}^n$ with density e^{-T} with respect to the Lebesgue measure $\mathrm{d}p$ (for example a Gaussian distribution, $T(p) \equiv \frac{1}{2} p^\top G^{-1} p + c$, for some inner product matrix G and normalising constant c, which corresponds to a left-invariant metric). Then the Hamiltonian should be defined as $H(g, p) = V(g) + T(p)$, since

$$\mu_H(q, p) = \frac{1}{Z_P} e^{-V(q)} e^{-T(p)} \theta^1 \wedge \cdots \wedge \theta^n \wedge \mathrm{d}p_1 \wedge \cdots \mathrm{d}p_n = \pi^* P \wedge \zeta.$$

[5] This may thought as the definition of ζ, and defines a differential form due to the transformation rules for tangent-lifted coordinates, see (2.5) [166]

3.4 The Hamiltonian Monte Carlo Algorithm

In §3.4.1 we will define the Hamiltonian Monte Carlo algorithm in its general form, which combines an appropriate numerical integration of the mechanics with a Metropolis algorithm to obtain an algorithm called MDMC that samples from the canonical distribution on an arbitrary symplectic manifold, which is further combined with a thermostat process to improve the ergodicity of the method. In §3.4.2 we will then consider various implementations of HMC based on full and partial momentum refreshment.

3.4.1 Sampling Canonical Distribution and Target Measure

Since we cannot compute the Hamiltonian flow exactly, we need to consider instead a numerical integrator $\Psi_{\delta t}$ of the Hamiltonian trajectory, which will in general no longer preserve the canonical distribution $\mu_H \propto e^{-H}\omega^n$. This can be rectified by adding a Metropolis-Hastings step whenever we can compose $\Psi_{\delta t}$ with a map $S : \mathcal{F} \to \mathcal{F}$ s.t., $(S \circ \Psi_{\delta t})^{-1} = S \circ \Psi_{\delta t}$, in which case $Q(z, \cdot) \equiv \delta_{S\circ\Psi_{\delta t}(z)}$ can be used as a proposal kernel and combined with an accept/reject step satisfying the conditions described in 2 [341] (here δ_z denotes the Dirac measure centred at z).

Let us now construct an explicit Markov transition kernel for HMC. Consider an **antisymplectic involution** $S : \mathcal{F} \to \mathcal{F}$ on a $2n$-dimensional symplectic manifold $(\mathcal{F}, \omega)$, i.e., $S^*\omega = -\omega$ and $S = S^{-1}$. A Hamiltonian H is said to have a **time-reversal (discrete) symmetry**, or simply to be **reversible** if $H \circ S = H$. Indeed in that case time-reflection $t \mapsto -t$ is obtained by conjugation $\Phi^H_{-t} = S \circ \Phi^H_t \circ S$, which follows from $\partial S(X_H) = -X_H \circ S$ [2, proposition 4.3.13]. Standard HMC Hamiltonians, being the sum of a Riemannian kinetic energy and a potential energy, are reversible with respect to the sign flip of the fibre component $S(q, v) = (q, -v)$, though, as we will see later, one has to be more careful when considering electromagnetic forms. Using $S^*\omega^n = (-1)^n\omega^n$, it follows that S preserves the one-density $|\omega^n| \equiv |\cdot| \circ \omega^n$, and thus the canonical distribution μ_H:

$$\int f S_\sharp \mu_H = \int f \circ S \mu_H = \int S^*(f) e^{-H}\omega^n = (-1)^n \int S^*\left(f e^{-H}\omega^n\right) = \int f \mu_H.$$

We will say a numerical integrator $\Psi_{\delta t}$ of the mechanics is **time-reversible** when $\Psi_{-\delta t} = S \circ \Psi_{\delta t} \circ S$, (time) **symmetric** when $\Psi_{-\delta t} = \Psi^{-1}_{\delta t}$,[6] and is volume-preserving if $(\Psi_{\delta t})_\sharp \omega^n = \omega^n$;

[6] The name "symmetric" comes from the fact that the **adjoint integrator** is defined as $\Psi^*_{\delta t} \equiv \Psi^{-1}_{-\delta t}$. Hence symmetric integrators are sometimes also referred to as Hermitian integrators. By theorem 1 [212] the symmetry of Ψ implies it must have even order. We mention as well that some authors call a numerical method **reversible** if $\Psi^{-1}_{\delta t} = S \circ \Psi_{\delta t} \circ S$.

S-Reversible Integrator	$\Psi_{\delta t}^{-1} = S \circ \Psi_{\delta t} \circ S$, where S bijection
Time-symmetric	$\Psi_{\delta t}^{-1} = \Psi_{-\delta t}$
Time-reversal symmetry	$S^*\omega = -\omega$, $S = S^{-1}$
Time-reversible	$\Psi_{-\delta t} = S \circ \Psi_{\delta t} \circ S$
Volume-preserving	$(\Psi_{\delta t})_\sharp \omega^n = \omega^n$
Reversible Hamiltonian	$S^* H = H$

Table 3.1: Properties of integrators

see table 3.1 for a summary of these properties. For example the leapfrog integrator, defined as $\Psi_{\delta t} \equiv \left(\Phi^V_{\frac{1}{2}\delta t} \circ \Phi^T_{\delta t} \circ \Phi^V_{\frac{1}{2}\delta t}\right)^{\tau/\delta t}$ (see §3.5.1), satisfies these properties since any Hamiltonian flow Φ^f_t is time-reversible, symplectic and symmetric (provided $S^* f = f$). Now, for a time-reversible, symmetric and volume-preserving integrator, and a reversible Hamiltonian, the time-homogeneous Markov chain given by the generalised Metropolis algorithm with proposal kernel $Q(z, \cdot) \equiv \delta_{\Psi_{\delta t}(z)}$, which has transition kernel

$$\mathbb{P}(z \to \mathrm{d}z^*) \equiv \min\left(1, r(z, z^*)\right) \delta_{\Psi_{\delta t}(z)}(\mathrm{d}z^*) + \left(1 - \min\left(1, r(z, \Psi_{\delta t}(z))\right)\right) \delta_{S(z)}(\mathrm{d}z^*),$$

preserves the canonical distribution μ_H, [7] and the Metropolis ratio simplifies to

$$r(z, z^*) = \exp\left(-(H(z^*) - H(z))\right)$$

as a result of the properties of $\Psi_{\delta t}$ (we will prove a far more general result in §4.3.3). The acceptance rate depends only on the energy difference of z and z^*, which means that if we are able to preserve energy over long trajectories the method will be able to generate distant proposals that favour decorrelation. Given $z^n \in \mathcal{F}$ the resulting algorithm is (see 2.1.4 [329])

- Let $z^* \equiv \Psi_{\delta t}(z^n)$ (i.e., sample z^* from $Q(z^n, \cdot)$)
- accept z^* with probability $\min\left(1, e^{-(H(z^*) - H(z^n))}\right)$. If accepted, then $z^{n+1} \equiv z^*$. Else $z^{n+1} \equiv S(z^*)$.

This is called the **molecular dynamics Markov chain** (MDMC) step. In order to have a good acceptance rate the numerical integrator is tuned so as to produce small energy changes during the MDMC step, in the sense that $H(z^n)$ and $H(z^{n+1})$ are close. However if the energy does not (or barely) changes during the MD step, the proposed sample will remain stuck in a energy level set $H = c$, see Fig. 3.1.

[7] This follows from $S_\sharp \mu_H = \mu_H$ which implies the transition kernel $\mathbb{P}$ is μ_H**-reversible up to** S, i.e., it satisfies the **detailed balance condition up to** S (or **modified detailed balance**) with respect to μ_H, $\mathbb{P}(z \to \mathrm{d}z^*)\mu_H(\mathrm{d}z) = \mathbb{P}(S(z^*) \to S(\mathrm{d}z))\mu_H(\mathrm{d}z^*)$, which in turn implies μ_H is an invariant distribution for the Markov chain (see 2.1.4.2 [329]).

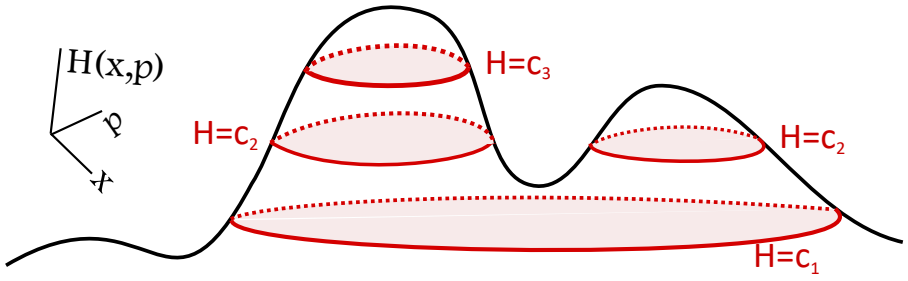

Figure 3.1: *Energy level sets*

Hence to obtain an ergodic chain it is necessary to combine MDMC with a μ_H-preserving momentum refreshment step (we shall discuss several such steps in what follows) called **heat bath or thermostat process**, which facilitates the exploration of different energy levels. The combination of MDMC with heat bath is known as **Hamiltonian Monte Carlo**: Given $z^n \in \mathcal{F}$

- Implement heat bath to obtain $\hat{z} \leftarrow z^n$
- Implement MD to propose $z^* \leftarrow \hat{z}$
- Implement MC to obtain $z^{n+1} \leftarrow z^*$.

If the aim is to sample from a target distribution P on $\mathcal{M}$, we construct a Hamiltonian system $(\mathcal{F}, \omega, H)$ adapted to P as explained in §3.3, so that the canonical distribution μ_H on the symplectic manifold $\pi : \mathcal{F} \to \mathcal{M}$ satisfies $\pi_\sharp \mu_H = P$. We then implement HMC to obtain samples (z^n) from μ_H, and project them to samples $(\pi(z^n))$ of P.

Note it is also common to include the involution within the proposal kernel, $Q(z, \cdot) \equiv \delta_{S \circ \Psi_{\delta t}(z)}$ (then $S \circ \Psi_{\delta t}$ is an involution) and use a standard Metropolis algorithm (with no involution S applied in case of rejection). When this MDMC is combined with a partial momentum refreshment (discussed in §3.4.2.2) it is then recommended to add a third step which flips the momentum in order to reduce the random walk behaviour (see 5.3 [276]).

3.4.2 Standard Implementations of HMC

We now discuss possible choices of heat bath and the resulting HMC schemes. We will begin by HMC with full momentum refreshment, §3.4.2.1, where we use the disintegration to sample a new momentum independently of its current value (this will motivate the general bracket MCMC considered later). In §3.4.2.2 we will generalise a scheme due to Horowitz which only partially refreshes the momentum, and §3.4.2.3 we will discuss an irreversible HMC algorithm constructed using a second order Langevin process.

3.4.2.1 *HMC with Full Momentum Refreshment*

Given a Hamiltonian system adapted to P with disintegration ζ as in §3.3, i.e., $\mu_H = e^{-H}\omega^n = \pi^* P \wedge \zeta$, HMC with full momentum refreshment uses ζ to perform the heat bath: given current phase $z = (q, p) \in \mathcal{F}$, the heat bath outputs $(q, \tilde{p})$ where $\tilde{p}$ is sampled from $\zeta|_{\pi^{-1}(q)}$. Note the distribution of $\tilde{p}$ is independent of p, and the heat bath is μ_H-preserving. We thus obtain 4.

Algorithm 4 Generate a sample from $P \propto e^{-V}\mu_{\mathcal{M}}$ given a current point $q \in \mathcal{M}$

1: **procedure** FULLREFRESHMENTHMC
2: Sample from the fibre $p \sim \zeta|_{\pi^{-1}(q)}$
3: Integrate Hamiltonian dynamics, $z^* \leftarrow \Psi_{\delta t}(q, p)$
4: Compute the Metropolis ratio $r \leftarrow e^{-(H(z^*)-H(q,p))}$
5: Accept the proposal $z \leftarrow z^*$ with probability $\min(1, r)$, else the new state is $z \leftarrow S(q, p)$
6: The P-sample is $q \leftarrow \pi(z)$

When $\mathcal{M}$ is a Riemannian manifold, and $P = \frac{1}{Z_P} e^{-V} \mathrm{vol}_{\mathcal{M}}$, we can consisder the Riemannian Newton system adapted to P as in §3.3.2. Sampling velocity according to $\zeta|_{\pi^{-1}(q)}$ corresponds to sampling from a Gaussian with mean 0 and covariance $\langle \cdot, \cdot \rangle_q^{-1}$. Since the Hamiltonian is $H = \pi^* V + \frac{1}{2}\| \cdot \|^2$, the involution $S(q, v) = (q, -v)$ does not affect the projected samples so may be ignored. In the following sections we focus on this Riemannian Newton Hamiltonian system adapted to P, and generalise the heat bath.

3.4.2.2 *HMC with Partial Momentum Refreshment*

In the HMC algorithm described above, at each step the momentum (or fibre element) is "fully refreshed" in the sense that the probability distribution of the momentum does not depend on its current value. In order to more accurately retain the dynamical properties and increase the acceptance rate (by setting the number of leapfrog steps to one without affecting the sample-correlation, thereby requiring less computational effort to propose new samples) [168] suggested replacing the momentum refreshment with a **partial momentum refreshment**. If we replace the step $v \sim \mathcal{N}(0, \langle \cdot, \cdot \rangle_q^{-1})$ in 4 with $v \sim \cos(\theta_q) v_0 + \sin(\theta_q)\mathcal{N}(0, \langle \cdot, \cdot \rangle_q^{-1})$ for some position-dependent mixing angle $q \mapsto \theta_q \in (0, \pi/2]$. Note that since the initial phase (q_0, v_0) is distributed according to $\propto e^{-\pi^* V - \frac{1}{2}\|\cdot\|_q^2}\omega_b^n$, then $v_0 \sim \mathcal{N}(0, \langle \cdot, \cdot \rangle_q^{-1})$, so $v \sim \cos(\theta_q) v_0 + \sin(\theta_q)\mathcal{N}(0, \langle \cdot, \cdot \rangle_q^{-1}) \sim \mathcal{N}(0, \langle \cdot, \cdot \rangle_q^{-1})$. The resulting method, called **generalised HMC** (GHMC), still preserves the target distribution. When $\theta = \frac{1}{2}\pi$ we recover standard HMC, while for arbitrary constant θ and a single leapfrog step we obtain the so-called **second order Langevin Monte Carlo** (L2MC) of [168]. However, while for a fixed step size L2MC does have a higher acceptance rate (due to the shorter trajectory length), and the partial momentum flip improves the mixing of the chain, the benefits are suppressed by the momentum flip introduced at each rejection which

almost reverse the trajectory, so the resulting samples are more likely to be more correlated [188] which reduces the mixing. L2MC thus only improves on HMC when the acceptance rate is sufficiently high.

Algorithm 5 GHMC Algorithm to generate a sample, given time-reversible, symmetric, volume-preserving integrator $\Psi_{\delta t}$ and mixing angles $q \mapsto \theta_q$

procedure STANDARD-GHMC
  Given current state $(q_0, v_0) \in T\mathcal{M}$
  sample $\tilde{v} \sim \mathcal{N}(0, \langle \cdot, \cdot \rangle_q^{-1})$
  $v \leftarrow \cos(\theta_q) v_0 + \sin(\theta_q)\tilde{v}$
  $(q^*, v^*) \leftarrow S \circ \Psi_{\delta t}(q_0, v)$
  Accept $z \leftarrow (q^*, v^*)$ with Metropolis step, else choose $z \leftarrow (q_0, v)$
  Flip velocity $\tilde{z} \leftarrow S \circ z$ to reverse trajectory upon rejection
  return $\tilde{z}$

One option to reduce this issue is to grant the algorithm $K \in \mathbb{N}$ **extra chances**: when the value z^* proposed by the numerical integrator should be rejected, we instead propose a new phase by integrating again the dynamics starting from z^* (this is done up to K times) (see Fig 3 [67]). Some numerical simulations are provided in [67] showing that with $K = 3$ extra chances, the effective sample size is substantially improved, in particular when θ is small. A similar idea is proposed in [321], where the strategy is to minimise the number of momentum flips (associated with rejections) by replacing them with further leapfrog steps, which leads to an algorithm with no accept-reject step. Another way to increase the acceptance rate is to couple GHMC with **shadow Hamiltonians** as discussed in §A.4.2.

3.4.2.3 Irreversible HMC with Ornstein-Uhlenbeck Heat Bath

The standard GHMC algorithm 5 combines a partial momentum refreshment, with a MDMC step whose transition kernel $\mathbb{P}$ is P-reversible up to S, which means it satisfies the modified detailed balance with respect to P, $\mathbb{P}(z \to \mathrm{d}z^*)P(\mathrm{d}z) = \mathbb{P}(S(z^*) \to S(\mathrm{d}z))P(\mathrm{d}z^*)$ (see §3.4.1). In V.B [114] it is shown that GHMC satisfies modified detailed balance. Recently there has been some growing interest in designing MCMC methods which do not satisfy detailed balance (i.e., are irreversible) as the chain (X_t) can benefit from a reduced asymptotic variance (describing the fluctuations of $\frac{1}{T}\int_0^T f(X_t)\mathrm{d}t$ around $\int f \mathrm{d}P$ for appropriate observables f) and a faster convergence speed to the target [275, 104, 103]. The idea is to consider Langevin processes which preserve the canonical distribution. These are then discretised by a splitting method into an energy-preserving Hamiltonian part, and a thermostat process which allows the system to move into different energy levels (which replaces the momentum refreshment). The Hamiltonian and thermostat processes are then composed to define a transition kernel. While both these processes do satisfy detailed balance, their composition does not, and we end up with an

irreversible chain.

Consider a Hamiltonian $H(q,p) \equiv \frac{1}{2}p^\top G^{-1}p + V(q)$. Given matrix-valued maps $\gamma, \sigma : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ and a standard Brownian motion W_t in $\mathbb{R}^n$ the **Langevin process**

$$\mathrm{d}Q_t = G^{-1}P_t\mathrm{d}t, \quad \mathrm{d}P_t = -\nabla V(Q_t)\mathrm{d}t - \gamma(Q_t)G^{-1}P_t\mathrm{d}t + \sigma(Q_t)\mathrm{d}W_t \tag{3.9}$$

preserves the canonical distribution $\propto e^{-H(q,p)}\mathrm{d}q\mathrm{d}p$ provided the fluctuation-dissipation relation holds: $\gamma = \frac{1}{2}\sigma\sigma^\top$ (see 2.2.3 [329]). A numerical integrator for (3.9) is formed by splitting it into a Hamiltonian part

$$\mathrm{d}Q_t = G^{-1}P_t\mathrm{d}t, \quad \mathrm{d}P_t = -\nabla V(Q_t)\mathrm{d}t,$$

and a **thermostat process**

$$\mathrm{d}Q_t = 0, \quad \mathrm{d}P_t = -\gamma(Q_t)G^{-1}P_t\mathrm{d}t + \sigma(Q_t)\mathrm{d}W_t.$$

The Hamiltonian part is approximated by a volume-preserving, time-reversible and symmetric integrator, coupled with a accept-reject step, while the thermostat part is a Ornstein-Uhlenbeck (OU) process which preserves the target measure and whose transition density can be found explicitly (see 5.1 [56]). Moreover in the limit in which the OU process is ran for an infinite time (i.e., we are sampling from its stationary distribution) then this method recovers HMC.

The irreversible HMC method described above can be extended to Hilbert spaces as explained in [289, 288] and to Lie groups [17]. See also [100] to the generalisation of the algorithm to distributions defined on the bounded hypercubes $[-a,a]^n$, rather than the whole $\mathbb{R}^n$. This may be achieved by either rejecting proposals that do not belong to the domain or using reflections at the boundary of the domain, though numerical simulations of the normalized effective sample size indicate that in high dimensions reflecting off the boundary performs better, and also outperforms HMC with reflections. Irreversible HMC can also be combined with shadow Hamiltonians, and the article [302] discusses the irreversibility resulting from the composition of momentum refreshment (which has an accept-reject step) with the MDMC step, and provides a nice summary of shadow HMC methods. We will discuss in §4 the geometrisation of the Langevin process as well as irreversible HMC schemes on manifolds.

We mention that such "metropolised-integrators" are also useful for SDEs in molecular dynamics. In that context one is usually interested in approximating accurately the SDE, and even though the accept-reject step (in particular the momentum flips) compromises the accuracy of numerical method, it can also improve the stability of the method as well as maintaining ergodicity with respect to the target distribution [56, 57]. On a similar note, the effect of the momentum flip in GHMC is analysed [7] where it is shown that on one hand momentum flips can be beneficial from the point of view of sampling, but on the other hand a modification of

GHMC with no momentum flip reproduces more accurately the molecular dynamics as long as the rejection rate is low (below 10% in their simulations).

3.5 Mechanical Integrators and Shadow Hamiltonians

As mentioned before, in general it is not possible to solve Hamilton's equations (3.7) and one must resort to approximate the exact solution with a numerical integrator $\Psi_{\delta t} : \mathcal{F} \to \mathcal{F}$. It is usually desirable to employ a **mechanical (or geometric) integrators**, that is one that preserves some of the geometric characteristics of the system, such as its energy, symplecticity, or symmetries (see 9 [250] for a discussion). For HMC we will be mainly interested in symplectic integrators, $\Psi^*_{\delta t}\omega = \omega$, as these will guarantee the acceptance rate remains high over long trajectories (see §3.5.4), even though symplecticity is a stronger property than required for HMC to be valid (recall from §3.4.1 that we only need volume-preservation). In the Hamiltonian framework, symplectic integrators are constructed through the splitting method, and in §3.5.1 we will discuss geodesic integrators, such as the one used in "geodesic Lagrangian Monte Carlo" [161] (which simply corresponds to leapfrog in velocity space). The splitting method, together with the fact that the target density is typically only known up to normalisation, also motivate the generalised bracket dynamics we consider in §4. The integrator proposed in [202] for "Lagrangian" Monte Carlo was non-symplectic, but this has nothing to do with the fact the Lagrangian perspective was (supposedly) employed. Indeed, Lagrangian integrators are always symplectic, and in the appendix we discuss the Lagrangian perspective to clarify this point. In §A.5.5 we review the discrete variational principle, which follows by discretising the principle of stationary action, from which we can derive the (unconstrained) discrete Euler–Lagrange equations (A.6) and the associated equations for the discrete Hamiltonian flow (A.7). Integrators obtained using the discrete variational principle, such as the generalised leapfrog method §A.5.5.1, are called **variational integrators**. Below (§3.5.2) we will consider the constrained discrete Euler-Lagrange equations (3.10) on manifolds with holonomic constraints, which are common in **molecular dynamics** as we usually have no closed-form expressions for the geodesics so cannot employ geodesic integrators. We will then discuss the theory of generating functions §3.5.3 which provides the general description of symplectic maps, and their Hamilton–Jacobi PDE (see §A.5.6), which may be used to prove the existence of a shadow Hamiltonian for symplectic integrators that controls the HMC acceptance rate as we discuss in §3.5.4 (and will also be useful for §4 when we discuss the shadow property of bracket mechanics).

3.5.1 Geodesic Integrators on $T\mathcal{M}$

Given a Hamiltonian $H = \sum_i H_i$, the **splitting method** offers a convenient procedure to construct symplectic integrators, and consists in approximating $X_H = \sum_i X_{H_i}$ by composing the flows of X_{H_i}. Consider a Riemannian Newton system adapted to P as in §3.3.2, with $T \equiv \frac{1}{2}\|\cdot\|^2$. If we are able to compute the geodesic flow Φ^T, it is then natural to "split" $H = T + V$ into Hamiltonians T and V in order to approximate Φ^H by a composition of geodesic Φ^T and vertical gradient Φ^V flows [212]. These mechanical integrators are referred to as **geodesic integrators** (see Fig. 3.2), and we will now introduce a particular one that belongs to the class of palindromic multi-stage integrators (see §2.1.3): the leapfrog integrator.

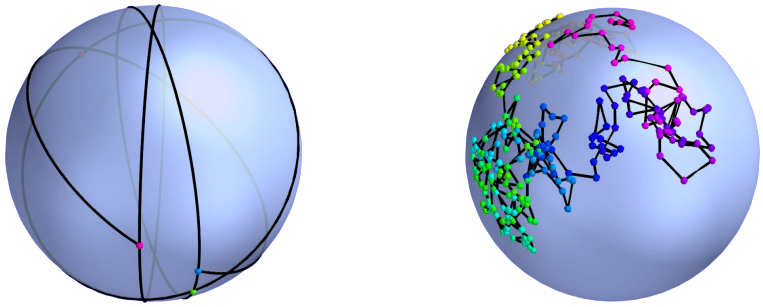

Figure 3.2: *A typical geodesic integrator alternates between geodesic steps, and vertical gradient steps that maintain the position but changes the direction. The figure illustrates the impact of the choice of numerical trajectory length.*

As mentioned above the solution $\dot\gamma : [a,b] \to T\mathcal{M}$ of (3.6) with Hamiltonian $H \equiv T$ (which is the integral curve of the Hamiltonian vector field X_T) defines the geodesic flow, i.e., $\gamma(t)$ is a geodesic from $\gamma(a)$ to $\gamma(b)$ and $\dot\gamma(b)$ is the parallel transport of $\dot\gamma(a)$. For $H = V \circ \pi$, we see from (3.6) that $X^i = 0$ and $\overline{X}^r = -G^{rs}\partial_{q^s} V$, which implies the vertical gradient update

$$q(t) = q(0) \qquad v(t) = v(0) - t\nabla V|_{q(0)}.$$

The **leapfrog (or Störmer/Verlet) integrator** with trajectory length τ and time step δt defines the following approximation of Φ^H

$$\Phi^H_\tau \approx \left(\Phi^V_{\frac{1}{2}\delta t} \circ \Phi^T_{\delta t} \circ \Phi^V_{\frac{1}{2}\delta t}\right)^{\tau/\delta t},$$

and may be implemented as follows 6

When the flow Φ^T is unknown, but we do have access to the geodesic flow $\tilde\Phi$ on $\mathcal{M}$ of some other Riemannian metric, it is possible to use $\tilde\Phi$ to derive a symplectic integrator as shown in [210]. The construction relies on generating functions §3.5.3, and reduces to the leapfrog

Algorithm 6 Leapfrog numerical method with $k \in \mathbb{N}_{\geq 1}$ trajectories of time step $\delta t > 0$.

procedure LEAPFROG
  Given current state $(q_0, v_0) \in T\mathcal{M}$
  $v_{\frac{1}{2}} \equiv v_0 - \frac{1}{2}\delta t \nabla V|_{q_0}$
  $(q_1, v_1^*) \equiv \Phi_{\delta t}^T(q_0, v_{\frac{1}{2}})$
  for for $m \in [1, k-1]$ **do**
    $v_{m+\frac{1}{2}} \equiv v_m^* - \delta t \nabla V|_{q_m}$
    $(q_{m+1}, v_{m+1}^*) \equiv \Phi_{\delta t}^T(q_m, v_{m+\frac{1}{2}})$
  $v_k \equiv v_k^* - \frac{1}{2}\delta t \nabla V|_{q_k}$
return (q_k, v_k)

integrator when the Riemannian metrics are proportional. In general the integrator is implicit. It considerably simplifies on symmetric spaces and compact Lie groups since these may be equipped with Riemannian metrics whose geodesic flow is particularly simple (see §3.1.3.1).

As mentioned earlier, it is worth noting that the splitting method only makes sense in the Hamiltonian framework. Indeed if we consider the potential energy as a Lagrangian $\mathcal{L} = V$ on Euclidean space, then (2.4) becomes the awkward $\partial_q V = 0$; while as we just saw $H = V$ implies $v(t) = v(0) - t\partial_q V$ and $\dot{q} = 0$, which shows the "decoupling" between the "velocity of the curve" $\dot{q}$, which must vanish, and the velocity coordinate v which varies linearly in $\partial_q V$. Indeed [252, theorem 1.4.3] shows $\mathcal{L}$ must be "regular" for the Lagrangian flow to be defined. As discussed in the appendix, the Lagrangian perspective instead relies on approximating the Lagrangian to discretise the action, and then use the discrete variational principle. Interestingly, the usual integrators for molecular simulations, RATTLE and SHAKE, are obtained via the constrained Euler–Lagrange equations as explained below. Hence, if there really is such a thing as Lagrangian Monte Carlo, it is really HMC for molecular systems with holonomic constraints.

3.5.2 HMC on Manifolds with Holonomic Constraints

Intuitively a mechanical system on $\mathbb{R}^n$ is said to have holonomic constraints if it is constrained to remain on an embedded submanifold $\mathcal{M}$ due to the presence of infinite force fields directed towards $\mathcal{M}$ which prevent particles from leaving $\mathcal{M}$ [19]. Holonomic constraints arise for example in directional statistics where distributions are defined on appropriate level sets, and in molecular systems where constraints are imposed, for example by fixing the bond length or bond angle of flexible molecules, in order to improve the computational efficiency by increasing the step-size at which the dynamics can be accurately simulated [349]. We say an embedded submanifold $\mathcal{M} \subset \mathbb{R}^n$ has (scleronomic) **holonomic constraints** if $\mathcal{M} \equiv \phi^{-1}(0)$ is the level

set $\phi : \mathbb{R}^n \to \mathbb{R}^m$ of a smooth map for which 0 is a regular value[8]. As a result $\mathcal{M}$ is a closed embedded submanifold of dimension $n - m$ [209, corollary 5.23], and each component ϕ_i may be viewed as an independent constraint (we assume $n > m$). The tangent bundle of $\mathcal{M}$ may then be identified with

$$T\mathcal{M} \cong \{(q, v) \in T\mathbb{R}^n : \phi(q) = 0, \partial\phi(v) = 0\}.$$

However, in general there is no canonical way to embed $T^*\mathcal{M}$ into $T^*\mathbb{R}^n$, and moreover the restriction of the symplectic structure from $T^*\mathbb{R}^n$ to the "primary constraint set" is usually degenerate (see 3.3.2 [252]). Fortunately, for a Lagrangian on $\mathcal{M}$ obtained (by restriction) from a hyperregular Langrangian on $\mathbb{R}^n$, the cotangent bundle may be (symplectically) embedded in $T^*\mathbb{R}^n$ and the mechanics on $\mathcal{M}$ may then be expressed as a constrained mechanics on $T\mathbb{R}^n|_{\mathcal{M}}$ [252]. Let us discuss this in more detail. Suppose the Lagrangian on $\mathcal{M}$ is obtained from the standard HMC Lagrangian $\mathcal{L}(q, v) \equiv \frac{1}{2}v^\top M(q)v - V(q)$ of $\mathbb{R}^n$. Then the symplectic embedding yields ($\sharp$ being the musical isomorphism)

$$T^*\mathcal{M} \cong \{(q, p) \in T^*\mathbb{R}^n : \phi(q) = 0, \partial\phi(p^\sharp) = 0\}.$$

We now relate mechanics on the submanifolds to mechanics on its embedding. Given a Lagrangian $\mathcal{L} : T\mathbb{R}^n \to \mathbb{R}$, its restriction $\mathcal{L}_{\mathcal{M}} \equiv \mathcal{L}|_{T\mathcal{M}} : T\mathcal{M} \to \mathbb{R}$ defines a mechanical system on $T\mathcal{M}$, and it can be shown that the Euler-Lagrange equations on $T\mathcal{M}$ (defined by $\mathcal{L}_{\mathcal{M}}$) are equivalent to a constrained Euler–Lagrange system on $\mathbb{R}^n$ for the Lagrangian $\mathcal{L}$, and to an unconstrained Euler–Lagrange system associated to an augmented Lagrangian on $\mathbb{R}^n \times \mathbb{R}^m$ [252, theorem 3.3.2]. In fact if $\partial\iota : T\mathcal{M} \to T\mathbb{R}^n$ denotes the tangent map of the inclusion $\iota : \mathcal{M} \to \mathbb{R}^n$, the solutions of the Euler-Lagrange system on $T\mathcal{M}$ are given by curves q on $\mathbb{R}^n$ which solves the projected Euler-Lagrange equations:

$$(\partial\iota)^* \left(\frac{\partial \mathcal{L}}{\partial q^i}(q, \dot{q}) - \frac{\mathrm{d}}{\mathrm{d}t} \left(\frac{\partial \mathcal{L}}{\partial v^i}(q, \dot{q}) \right) \right) = 0.$$

With these definitions we can now derive the constrained discrete Euler–Lagrange equations. Given a Lagrangian $\mathcal{L}(q, v) \equiv \frac{1}{2}v^\top M(q)v - V(q)$ on $T\mathbb{R}^n$, approximated by a discrete Lagrangian $\mathcal{L}_d : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$. We restrict these to the embedded submanifold and obtain $\mathcal{L}_{\mathcal{M}}$ and $\mathcal{L}_d^M : \mathcal{M} \times \mathcal{M} \times \mathbb{R} \to \mathbb{R}$. Fix $N \in \mathbb{N}$ and endpoints $q_0, q_{Nh} \in \mathcal{M}$. We then have the following by theorems 3.4.1-3.4.2 [252] : a discrete curve $(q_0, q_1, \ldots, q_{Nh}) \in \mathcal{M}^{N+1}$ extremises the discrete action $\mathcal{S}_d$ on $\mathcal{M}$, and thus is a solution of discrete Euler-Lagrange equations on $\mathcal{M}$, iff $(q_0, q_1, \ldots, q_{Nh}) \in (\mathbb{R}^n)^{N+1}$ (viewed as a discrete curve in $\mathbb{R}^n$) and the Lagrange multipliers

[8]Recall this means $\partial_q\phi$ is surjective for all $q \in \phi^{-1}(0)$, i.e., $\phi : \phi^{-1}(0) \to \mathbb{R}^m$ is a submersion

$(\lambda_1, \ldots, \lambda_{N-1}) \in (\mathbb{R}^n)^{N-1}$ solve the **constrained discrete Euler–Lagrange equations**

$$\nabla_2 \mathcal{L}_d(q_{j-1}, q_j) + \nabla_1 \mathcal{L}_d(q_j, q_{j+1}) = \lambda_k^\top \nabla \phi(q_j) \tag{3.10a}$$

$$\phi(q_j) = 0. \tag{3.10b}$$

Moreover the symplectic integrator $(q_k, p_k) \mapsto (q_{k+1}, p_{k+1})$ generated by the associated discrete Hamiltonian flow is $\Phi^H_{\mathcal{L}_d^{\mathcal{M}}} : T^*\mathcal{M} \to T^*\mathcal{M}$ (see §A.5.5),

$$p_k = -\nabla_1 \mathcal{L}_d(q_k, q_{k+1}) + \bar{\lambda}^\top \nabla \phi(q_k) \tag{3.11a}$$

$$p_{k+1} = \nabla_2 \mathcal{L}_d(q_k, q_{k+1}) - \dot{\lambda}^\top \nabla \phi(q_{k+1}) \tag{3.11b}$$

$$0 = \phi(q_{k_1}) \tag{3.11c}$$

$$0 = \nabla \phi(q_{k+1}) M^{-1}(q_{k+1}) p_{k+1} \tag{3.11d}$$

where $\bar{\lambda}, \dot{\lambda} \in \mathbb{R}^m$ are Lagrange multipliers.

3.5.2.1 SHAKE, RATTLE and HMC on Level Sets

As discussed in the appendix §A.5.5.1, the leapfrog integrator for the Lagrangian $\mathcal{L} = \frac{1}{2} v^\top M v - V(q)$ can be obtained from a discrete Lagrangian given by a trapezoidal rule. The solution of (3.10) defines the SHAKE algorithm

$$M\left(\frac{q_{k+1} - 2q_k + q_{k-1}}{2}\right) + h\nabla V(q_k) + \lambda_k^\top \nabla \phi(q_k) = 0,$$

$$\phi(q_{k+1}) = 0.$$

The associated symplectic integrator over $T^*\mathcal{M}$ is called RATTLE, and can be found using (3.11), which after some re-arranging and rescaling yields

$$\begin{aligned}
q_{k+1} &= q_k + hM^{-1}p_k - \tfrac{1}{2}h^2 \nabla V(q_k) + \tfrac{1}{2}h^2 M^{-1} \bar{\lambda}_k^\top \nabla \phi(q_k) \\
p_{k+1} &= p_k - h\left(\frac{\nabla V(q_k) + \nabla V(q_{k+1})}{2}\right) + h\left(\frac{\bar{\lambda}_k^\top \nabla \phi(q_k) + \dot{\lambda}_k^\top \nabla \phi(q_{k+1})}{2}\right) \\
0 &= \phi(q_{k+1}) \\
0 &= \nabla \phi(q_{k+1}) M^{-1} p_{k+1}.
\end{aligned}$$

In practice RATTLE is often used in the following format:

$$p_{k+\frac{1}{2}} = p_k - \tfrac{1}{2}h\nabla V(q_k) + \tfrac{1}{2}h\bar{\lambda}_k^\top \nabla \phi(q_k) \tag{3.12a}$$

$$q_{k+1} = q_k + hM^{-1}p_{k+\frac{1}{2}} \tag{3.12b}$$

$$0 = \phi(q_{k+1}) \qquad \text{(project onto configuration manifold)} \tag{3.12c}$$

$$p_{k+1} = p_{k+\frac{1}{2}} - \tfrac{1}{2}h\nabla V(q_{k+1}) + \tfrac{1}{2}h\dot{\lambda}_k^\top \nabla\phi(q_{k+1}) \tag{3.12d}$$

$$0 = \nabla\phi(q_{k+1})M^{-1}p_{k+1} \qquad \text{(project onto cotangent space).} \tag{3.12e}$$

By construction the RATTLE numerical integrator (3.12) is symplectic. It is also symmetric (see theorem 1.3 chapter VII [148]) and time-reversible (see 7.7 [212]). Hence it can be used as the numerical integrator of a HMC algorithm to sample distributions on manifolds defined as a zero level set $\mathcal{M} \equiv \phi^{-1}(0)$, provided the target density is expressed as $P \propto e^{-V}\text{vol}$ where $V : \mathbb{R}^n \to \mathbb{R}$ is a function defined on the embedded space and vol the Riemannian volume form induced by restricting the constant Riemannian metric M to $\mathcal{M}$. This was applied to statistics in [63]. However, the reversibility of RATTLE only actually holds for sufficiently small time steps, and in practice it is necessary to include a reversibility-check when the the time-step is not sufficiently small in order to sample without bias, as explained in [214]. Hence, to sample without bias with the RATTLE integrator, one should include the reversibility check proposed in [214, 367]. Recently [211] argued that instead of using SHAKE/RATTLE to implement the constrained dynamics, it was preferable to first split the Hamiltonian into the usual sum of the kinetic and potential energies, and use a composition method which computes the potential and kinetic flow separately and composes them. The constrained mechanics corresponding to the potential energy can be explicitly obtained, while the free motion (kinetic flow) can for example be computed using SHAKE/RATTLE (see the appendix of [211] for the detailed algorithm). Using this composition method led to a substantial improvement of the computational stability and accuracy which permitted the use of larger step-sizes, at the cost of biased samples. Of course one can also combine a geodesic integrator with the RATTLE with reversibility check integrator only applied on the intractable geodesic motion - see [139] for some applications to partially observed diffusion models. See also [151] for a discussion of ergodicity on holonomic systems, and [259] for generalisations of SHAKE/RATTLE.

While we shall only apply the theory of symmetries we discuss later to geodesic orbit manifolds, it would be interesting in the future to use it to construct new numerical integrators outperforming SHAKE/RATTLE for holonomic systems. One option we have been discussing with Professor Lelièvre, is to construct a Hamiltonian system with symmetries in the embedding space whose mechanics remains on the constrained system without Lagrange multipliers. A natural place to start is to combine the Impetus-Striction formulation of mechanics [356], which has the desired properties, together with the explicit integrator for non-separable Hamiltonian recently constructed in [335].

3.5.3 Generating Functions and Symplectic Integrators

Splitting and variational integrators are symplectic, and thus may also be obtained using the more general theory of generating functions. The method of generating functions is built on the fact that a map

$$F : T^*\mathcal{M} \to T^*\mathcal{M} \quad \text{is symplectic iff} \quad \mathrm{d}\iota_F^*\Xi = 0$$

where $\Xi \equiv \pi_1^*\Theta - \pi_2^*\Theta$ with Θ the canonical 1-form, $\pi_i : T^*\mathcal{M} \times T^*\mathcal{M} \to T^*\mathcal{M}$ are the standard projections onto components, and $\iota_F : \Gamma(F) \hookrightarrow T^*\mathcal{M} \times T^*\mathcal{M}$ is the inclusion of the graph of F (proposition 5.2.1 [2]). By Poincaré Lemma, since the 1-form $\iota_F^*\Xi$ defined on the graph $\Gamma(F)$ is closed, there exists locally a map $S : \Gamma(F) \to \mathbb{R}$, called the **generating function** of F.

Suppose now that we are given a generating function for F and want to reconstruct F. This is first achieved by choosing local coordinates on $\Gamma(F)$. If (q, p, Q, P) are coordinates on $T^*\mathcal{M} \times T^*\mathcal{M}$, then on $\Gamma(F)$ we can relate these coordinates by $(Q(q,p), P(q,p)) \equiv F^*(Q,P)|_{(q,p)} = (Q \circ F, P \circ F)|_{(q,p)}$, and in particular (q, p) define local coordinates on $\Gamma(F)$. Typically however S is given as a function $S(q, Q)$, $S(p, Q)$ or $S(P, q)$, i.e., we proceed to a change of coordinates on $\Gamma(F)$. [9] If we choose the local coordinates (q, Q) on $\Gamma(F)$ then $\iota_F^*\Xi = \mathrm{d}S$ reads

$$p = -\frac{\partial S}{\partial q}(q, Q), \qquad P = \frac{\partial S}{\partial Q}(q, Q). \tag{3.13}$$

These equations implicitly define the symplectic map $F : (q, p) \mapsto (Q, P)$. From above we also have that a function $S : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ generates by (3.13) (when a solution exists) a symplectic map $(q, p) \mapsto (Q, P)$.

3.5.3.1 Discrete Lagrangians as Generators

Fix a time step δt and consider a discrete Lagrangian $\mathcal{L}_d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$. As discussed in the appendix, $\mathcal{L}_d$ defines a symplectic integrator by (A.7)

$$p_j = -\partial_1\mathcal{L}_d(q_j, q_{j+1,\delta t}), \qquad p_{j+1} = \partial_2\mathcal{L}_d(q_j, q_{j+1}, \delta t)$$

which are precisely the generating equations (3.13) for $S \equiv \mathcal{L}_d$. This shows that the discrete Lagrangian is precisely the generating function for the discrete Hamiltonian flow $\Phi^H_{\mathcal{L}_d}$.

[9] Recall that if we have a map $(q, p) \mapsto (Q(q,p), P(q,p))$, by the inverse function theorem we can express it as a function of the coordinates (q, Q) when the Jacobian of the change of coordinates $(q, p) \mapsto (q, Q)$ is a linear isomorphism, i.e., $\nabla(q, Q(q,p))$ is invertible, which holds whenever $\partial_p Q$ is invertible.

3.5.3.2 *Geodesic Flows as Symplectic Maps Generated by the Riemann Distance*

Let $\mathcal{M}$ be a geodesically convex Riemannian manifold, which means that any two points are connected by a unique geodesic whose length is the Riemann distance $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$. We can apply the method of generating function with $S \equiv -\frac{1}{2}d$ and use the symplectomorphism $T^*\mathcal{M} \to T\mathcal{M}$ given by the musical isomorphism (see §3.2.3) to obtain an associated symplectic map $T\mathcal{M} \to T\mathcal{M}$. It is shown in 4.3 [93] that by solving (3.13) one obtains the geodesic flow.

3.5.4 Modified Equations and Shadow Hamiltonians

We are finally ready to introduce precisely one of the key reasons for the success of HMC with symplectic integrators: the existence of a shadow Hamiltonian. The shadow Hamiltonian does not imply that the numerical trajectory is anywhere close to the original one (in fact there is no concept of distance on symplectic manifolds), but instead provides some energy conservation properties, which is precisely the quantity of interest to obtain a high acceptance rate (recall §3.4.1). We will define the notion of "modified equations" in §3.5.4.1, and the result about the existence of the shadow Hamiltonian, as typically presented on Euclidean space. In §3.5.4.2 we will discuss its Poisson bracket expansion, which may be used to tune integrators in extensive systems, and finally explain the role played by the Shadow Hamiltonian for the conservation of acceptance rate in §3.5.4.3.

3.5.4.1 *Modified Equations*

Consider $\mathcal{M} = \mathbb{R}^n$ so that phase space $T^*\mathbb{R}^n$ is a metric space. Recall the solution to Hamilton's equations with initial value $z_0 \in T^*\mathbb{R}^n$ is $z(t) = \Phi_t^H(z_0)$. As we discussed HMC requires a volume-preserving integrator $\Psi_{\delta t}$ with a given step size δt, which defines a numerical trajectory $z_{k+1} \equiv \Psi_{\delta t}(z_k) = \Psi_{\delta t}^{k+1}(z_0)$. The essential idea of backward error analysis is to find a field $f_\Psi : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ s.t., the **modified differential equation**

$$\dot{z} = f_\Psi(z)$$

is a better approximation to the numerical trajectory than the original trajectory. Using the Hamilton-Jacobi PDE (see §A.5.6) it can in fact be shown that for any symplectic method $\Psi_{\delta t}$ (consistent with X_H), there exists a formal **Shadow Hamiltonian** $\mathcal{H}_{\Psi_{\delta t}} \equiv H + \delta t H_1 + \delta t^2 H_2 + \delta t^3 H_3 + \cdots$, where each H_i is defined on the same domain as the generating function, whose Hamilton's equations are the modified equations of $\Psi_{\delta t}$ i.e., $z_k = \Psi_{\delta t}^k(z_0) = \Phi_{k\delta t}^{\mathcal{H}_{\Psi_{\delta t}}}(z_0)$. Combining results from [148, 34] we have

Theorem 1. *Suppose $\Psi_{\delta t}$ has a generating function $S(P,q) = \delta t S_1(P,q) + \delta t^2 S_2(P,q) + \cdots$ where each S_j is defined on an open set $D \subset \mathbb{R}^n \times \mathbb{R}^n$. Then the modified equations for $\Psi_{\delta t}$ are the Hamilton's equations of $\mathcal{H}_{\Psi_{\delta t}} \equiv H + \delta t H_1 + \delta t^2 H_2 + \cdots$ where H_j are smooth functions defined on D. Moreover we have for some $a, b > 0$*

$$\|\Phi_1^{\mathcal{H}_{\Psi_{\delta t}}} - \Psi_{\delta t}\| \leq a\, \delta t\, e^{-b/\delta t},$$

or in other words, the flow of the shadow Hamiltonian differs from the symplectic integrator for the flow of H by an "exponentially small amount", where "small" is determined by the ratio $b/\delta t$.

The proof is based on substituting the formal expansion of $\mathcal{H}_{\Psi_{\delta t}}$ into Hamilton-Jacobi PDE, and choosing the functions H_j to ensure its generating function is indeed S when $t \equiv \delta t$. From the proof we obtain a recurrence relation which enables us to compute each H_j from the generating function S.

3.5.4.2 Poisson Bracket Expansion of Shadow Hamiltonian

To make sense of the Shadow Hamiltonian expansion on manifolds we need to view vector fields as derivations[10] on the Lie algebra of functions $C^\infty(T^*\mathcal{M})$, $X \in \Gamma(TT^*\mathcal{M}) : (f \mapsto (p \mapsto X_p f)) \in \mathrm{Der}\,(C^\infty(T^*\mathcal{M}))$. Derivations form a Lie algebra with the commutator as the product. For $X \in \mathrm{Der}\,(C^\infty(T^*\mathcal{M}))$, we can define the formal exponential in the variable t

$$e^{tX} \equiv \mathrm{Id} + tX + \frac{t^2}{2!}X^2 + \cdots \in \mathrm{End}(C^\infty(T^*\mathcal{M}))[t].$$

We identify the flow Φ_t^H with the formal exponential $e^{t\hat{H}}$, since the latter acts as the pullback of the former $e^{t\hat{H}} f = f \circ \Phi_t^H$ (4.4 [116]), [11] from which we can check the exponential is indeed a solution, i.e., an integral curve of X_H

$$\frac{\mathrm{d}e^{t\hat{H}}}{\mathrm{d}t} f = e^{t\hat{H}} \hat{H}[f] = \hat{H}[f] \circ \Phi_t^H = \hat{H}_{\Phi_t^H}[f].$$

We can now use the correspondence between flows and exponential series to determine the shadow Hamiltonian of an integrator by splitting. By the Baker-Campbell-Hausdorff formula

[10] Recall a derivation on an algebra $\mathcal{A}$ is a linear map $d : \mathcal{A} \to \mathcal{A}$ s.t. $d(ab) = d(a)b + ad(b)$

[11] More generally, on tensor fields $e^{t\mathcal{L}_{\hat{H}}} = (\Phi_t^H)^*$. In Euclidean space, if $\frac{\mathrm{d}\gamma}{\mathrm{d}t} = \hat{H}(\gamma(t)) = \hat{H}^i \partial_{z^i}|_{\gamma(t)}$ then for any scalar function f, using a Taylor expansion of $f \circ \gamma$ around $t = 0$ we can show (see III [148]) $f \circ \Phi_t^H = e^{t\hat{H}^i \partial_{z^i}|_{\gamma(t)}} f$

(see theorem 6 [85] for definition of c_m)

$$\Phi_t^A \circ \Phi_s^B = e^{\hat{A}t} e^{\hat{B}s} = e^{\sum_{m=1} c_m(t\hat{A}, s\hat{B})} = e^{t\hat{A}+s\hat{B}+\frac{ts}{2}[\hat{A},\hat{B}]+\frac{t^2 s}{12}\left[\hat{A},[\hat{A},\hat{B}]\right]+\frac{ts^2}{12}\left[\hat{B},[\hat{B},\hat{A}]\right]+\cdots}.$$

Since the sum and commutator of Hamiltonian vector fields is a Hamiltonian vector field, [12] we can write the RHS as the flow of a **shadow Hamiltonian vector field** $\hat{\mathcal{H}}_t$, from which we can derive the Poisson bracket expansion of the shadow Hamiltonian $\mathcal{H}_t$, i.e., $\Phi_t^A \circ \Phi_t^B = e^{t\hat{A}} e^{t\hat{B}} = e^{t\hat{\mathcal{H}}_t} = \Phi_t^{\mathcal{H}_t}$, with $\mathcal{H}_t = A + B + \frac{1}{2}t\{A, B\} + \frac{1}{12}t^2\left(\{A, \{A, B\}\} - \{B, \{A, B\}\}\right) + \cdots$. Naturally this is a formal expansion, we refer to [187] for a discussion of its convergence.

We can apply this to the leapfrog integrator on Lie groups. For a Hamiltonian $H(g, X) = T(X) + V(g)$ on $\mathcal{G} \times \mathfrak{g}$, the shadow Hamiltonian of the leapfrog integrator is $\Phi^V_{\frac{1}{2}\delta t} \circ \Phi^T_{\delta t} \circ \Phi^V_{\frac{1}{2}\delta t}$

$$\mathcal{H}_{\delta t} = T + V - \frac{\delta t^2}{24}\left(\{V, \{V, T\}\} + 2\{T, \{V, T\}\}\right) + O(\delta t^4).$$

For the more general **two-stage integrator**, $\Phi^V_{\lambda\delta t} \circ \Phi^T_{\frac{1}{2}\delta t} \circ \Phi^V_{(1-2\lambda)\delta t} \circ \Phi^T_{\frac{1}{2}\delta t} \circ \Phi^V_{\lambda\delta t}$, then [284]

$$\mathcal{H}_{\delta t} = H + \delta t^2 \left(\frac{6\lambda^2 - 6\lambda + 1}{12}\{V, \{V, T\}\} + \frac{1 - 6\lambda}{24}\{T, \{V, T\}\}\right) + O(\delta t^4). \tag{3.14}$$

Another important example is the standard Hamiltonian $H(q, p) = \frac{1}{2}p^\top G^{-1} p + V(q)$ on Euclidean space $\mathbb{R}^n \times \mathbb{R}^n$, in which case for the **three-stage integrator** $\Phi^T_{b\delta t} \circ \Phi^V_{a\delta t} \circ \Phi^T_{(1-\frac{1}{2}b)\delta t} \circ \Phi^V_{(1-2a)\delta t} \circ \Phi^T_{(\frac{1}{2}-b)\delta t} \circ \Phi^V_{a\delta t} \circ \Phi^T_{b\delta t}$ with parameters $a, b \in \mathbb{R}$, the associated shadow Hamiltonian was derived in 3.1 [303]

$$\mathcal{H}_{\delta t} = H + \delta t^2 \left(c_{22} \nabla V^\top G^{-1} \nabla V + c_{21} p^\top G^{-1} \nabla^2 V G^{-1} p\right) + \mathcal{O}(\delta t^4)$$

with

$$c_{21} = \frac{1}{12}\left(1 - 6a(1-a)(1-2b)\right), \quad c_{22} = \frac{1}{24}\left(6a(1-2b)^2 - 1\right).$$

3.5.4.3 Conservation of Acceptance Rate

For numerical methods of order r (with sufficiently small δt), it can be shown the difference between the Shadow Hamiltonians and the true Hamiltonian is (5.1.2. [212])

$$|\mathcal{H}_{\Psi_{\delta t}} - H| = O(\delta t^r),$$

[12] Specifically: $\hat{A} + \hat{B} = \widehat{A + B}$, $[\hat{A}, \hat{B}] = \widehat{\{A, B\}}$

and that $\mathcal{H}_{\Psi_{\delta t}}$ is conserved over an exponentially large number of steps, i.e., there exists $a, b > 0$ s.t., (see equations (1.1-1.7) [34])

$$|\mathcal{H}_{\Psi_{\delta t}} \circ \Psi_{\delta t}^k - \mathcal{H}_{\Psi_{\delta t}}| \leq ak\delta t e^{-b/\delta t}.$$

Together, these imply that symplectic integrators preserve the Hamiltonian H up to order δt^r for exponentially long time. For HMC this has the important consequence that the acceptance rate remains high for long times. Unlike what is sometimes believed, this does not hold because the shadow trajectory is a good approximation of the true Hamiltonian trajectory, but rather despite the fact that the numerical trajectory typically "diverges" from the true one, yet remain close in "energy distance", which ensures that proposals with large integration time are often accepted, thus reducing the random walk behaviour and improving the mixing of the chain.

3.6 Hamiltonian Monte Carlo on Embedded Manifolds

We now consider the problem of sampling a distribution P on manifold $\mathcal{M}$, assuming we have an **embedding** $\mathcal{I} : \mathcal{M} \to \mathcal{E}$ into a "nice" manifold $\mathcal{E}$. This means that **(i)** $\mathcal{I}$ is injective, **(ii)** $\partial_p \mathcal{I}$ is injective for each $p \in \mathcal{M}$ (it is an immersion), and **(iii)** $\mathcal{I}$ is a homeomorphism onto its image (where $\mathcal{I}(\mathcal{M})$ is equipped with the subspace topology from $\mathcal{E}$). The image $\mathcal{I}(\mathcal{M})$ is then an embedded submanifold of $\mathcal{E}$. Expressing the HMC scheme on $\mathcal{M}$ in terms of its embedding $\mathcal{E}$ can be useful when (for example) $\mathcal{E}$ is flat or has global coordinates. The two most important cases consist of embedded submanifold of the vector space $\mathcal{E} \equiv \mathbb{R}^n$, and matrix subgroups of $\mathcal{E} \equiv \mathrm{GL}(n)$. The important case of embedded submanifolds of $\mathbb{R}^n$ defined by holonomic constraints was discussed in §3.5, but in this section we will assume we have access to the geodesic flow (as is common in statistical problems) so we can use the explicit leapfrog integrator §3.5.1 (or a more general multi-stage integrator §2.1.3) and can avoid the use of Lagrange multipliers. Hence our aim is to express the vertical gradient and geodesic steps of leapfrog in the embedding.

In §3.6.1 we first focus in the case where $\mathcal{I}$ is an isometry, and discuss the more general case in §3.6.2. We will then look at three important applications of the theory. In §3.6.3 we discuss matrix subgroups of the general linear group of invertible matrices $\mathrm{GL}(n)$; in §3.6.4 we consider the setting of [66] where $\mathcal{M}$ is an embedded submanifold of $\mathbb{R}^n$ equipped with the Euclidean metric, and finally in §3.6.5 we consider open submanifolds of vector spaces. This shows that the original HMC scheme was already an embedded scheme (into the curved manifold of invertible matrices) where the Lie algebra projection is a left-trivialised version

of the projection onto the tangent spaces, though the totally geodesic embedding of matrix groups makes its implementation simpler. It also shows that the Jacobian determinant of the Hausdorff density should never appear, in contrast with [66], but also to [160] according to which the presence of the Jacobian is due to a lack of symplecticity of geodesic Lagrangian Monte Carlo. Finally, it shows that the HMC method on covariance matrices developed in [161] is an embedded extrinsic scheme: the reason why there is no need to project onto the tangent spaces is simply due to the fact that the submanifold is open in the subspace topology.

3.6.1 Isometries are Symplectomorphisms

Let $\mathcal{I} : (\mathcal{M}, \langle\cdot,\cdot\rangle_{\mathcal{M}}) \to (\mathcal{N}, \langle\cdot,\cdot\rangle_{\mathcal{N}})$ be an isometric diffeomorphism (or isometry) between Riemannian manifolds, which means $\partial_p\mathcal{I}$ is an isometry for every $p \in \mathcal{M}$. Consider a Hamiltonian $H \equiv V \circ \pi + \frac{1}{2}\|\cdot\|^2 : T\mathcal{N} \to \mathbb{R}$. Then the tangent map $\partial\mathcal{I}$ is a symplectomorphism (by definition of $\omega_\flat$), and the mechanics with Hamiltonian H on $T\mathcal{N}$ is equivalent to the mechanics with Hamiltonian $H \circ \partial\mathcal{I}$ on $T\mathcal{M}$. In particular, the geodesic flow on $\mathcal{M}$ is the image under $\partial\mathcal{I}$ of the geodesics flow on $\mathcal{N}$,[13] and the Riemannian gradient of V and its pullback $W \equiv \mathcal{I}^*V$ to $\mathcal{M}$ are related by

$$\partial\mathcal{I}(\nabla^{\mathcal{M}}W) = \nabla^{\mathcal{N}}V.$$

Proof. Let $\mathcal{I} : \mathcal{M} \to \mathcal{N}$ be an isometry. For any $w = \partial\mathcal{I}(v) \in T\mathcal{N}$, then

$$\begin{aligned}\left\langle \partial\mathcal{I}(\nabla^{\mathcal{M}}W), w\right\rangle_{\mathcal{N}} = \left\langle \partial\mathcal{I}(\nabla^{\mathcal{M}}W), \partial\mathcal{I}(v)\right\rangle_{\mathcal{N}} = \left\langle \nabla^{\mathcal{M}}W, v\right\rangle_{\mathcal{M}} = \left\langle \nabla^{\mathcal{M}}(V\circ\mathcal{I}), v\right\rangle_{\mathcal{M}} = \\ = \mathrm{d}(V\circ\mathcal{I})(v) = \mathrm{d}U(w) = \left\langle \nabla^{\mathcal{N}}V, w\right\rangle_{\mathcal{N}}.\end{aligned}$$

■

We can thus perform HMC on either tangent bundles and then use $\partial\mathcal{I}$ to transport the samples [255] generated by the canonical distributions. In particular if we implement HMC on $\mathcal{N}$ to sample from $\tilde{P} \propto e^{-V}\mathrm{vol}_{\mathcal{N}}$, then the transported samples to $\mathcal{M}$ using $\mathcal{I}^{-1}$ will be distributed according to $P \propto e^{-W}\mathrm{vol}_{\mathcal{M}}$ (here vol is the Riemannian measure §A.5.1). Let us briefly mention an important isometry in statistics. The open simplex

$$\Delta^n_+ \equiv \{(x_1,\dots,x_{n+1}) \in \mathbb{R}^{n+1} : \sum x_i = 1, x_i > 0\}$$

[13] For $\mathcal{I}$ to map geodesics to geodesic it is in fact only necessary that it is a local isometry, i.e., $\partial_p\mathcal{I}$ is an isometry for all p

is a manifold whose tangent space at $x \in \Delta_+^n$ is

$$T_x\Delta_+^n = \{x\} \times \{v \in \mathbb{R}^{n+1} : \sum v_i = 0\}.$$

Discrete statistical models are subsets of Δ_+^n, and we will analyse in detail their continuous analogue discussed in §5.1. The open simplex is diffeomorphic to the positive quadrant of the hypersphere

$$S_+^n \equiv \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} : \sum x_i^2 = 1, x_i > 0\}$$

with diffeomprhism $\mathcal{I} : \Delta_+^n \to S_+^n$, $\mathcal{I}(x_1, \ldots, x_{n+1}) \equiv (\sqrt{x_1}, \ldots, \sqrt{x_{n+1}})$. We may turn $\mathcal{I}$ into an isometry by defining on Δ_+^n the pullback Riemannian metric from S_+^n (with standard metric), which yields the Fisher information metric $\langle \cdot, \cdot \rangle^F$ defined by

$$\langle u, v \rangle_x^F \equiv \sum \frac{u_i v_i}{x_i}.$$

Hence the multinomial manifold $(\Delta_+^n, \langle \cdot, \cdot \rangle^F)$ is isometric to S_+^n. We refer to 5.2 [66] for an application of HMC using this isometry to sample non-conjugate models on the simplex. We note that [66] considers the (closed) simplex $\Delta^n \equiv \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} : \sum x_i = 1, x_i \geq 0\}$ but the Fisher information metric is then not well-defined as it may involve a division by 0. See 2.2 [21] for a discussion on the extension of the Fisher metric to the closed simplex.

3.6.2 HMC via Embeddings

Suppose now that $\mathcal{I} : \mathcal{M} \to \mathcal{E}$ is an embedding. Then $\mathcal{I}(\mathcal{M})$ is a submanifold, and we denote the inclusion $\iota : \mathcal{I}(\mathcal{M}) \to \mathcal{E}$. Choose a nice metric $\langle \cdot, \cdot \rangle$ on $\mathcal{E}$ (i.e., one for which we can calculate the associated geodesics on $\mathcal{I}(\mathcal{M})$), which in turn defines a metric $\langle \cdot, \cdot \rangle^{\mathcal{I}(\mathcal{M})} \equiv \iota^* \langle \cdot, \cdot \rangle$ on $\mathcal{I}(\mathcal{M})$ by restriction, and a metric $\langle \cdot, \cdot \rangle^{\mathcal{M}}$ on $\mathcal{M}$ by setting $\mathcal{I}$ to be an isometry:

$$\langle v, w \rangle_q^{\mathcal{M}} = \langle \partial_q \mathcal{I}(v), \partial_q \mathcal{I}(w) \rangle_x^{\mathcal{I}(\mathcal{M})} \qquad \forall v, w \in T_q\mathcal{M}$$

where $x = \mathcal{I}(q)$. Hence $(\mathcal{M}, \langle \cdot, \cdot \rangle^{\mathcal{M}})$ and $(\mathcal{I}(\mathcal{M}), \langle \cdot, \cdot \rangle^{\mathcal{I}(\mathcal{M})})$ are isometric, and from §3.6.1 we know the mechanics on $T\mathcal{M}$ with energy $H(q, v) = V(q) + \frac{1}{2} \langle v, v \rangle^{\mathcal{M}}$ is equivalent to the mechanics on $\mathcal{I}(\mathcal{M}) \subset \mathcal{E}$ defined by the energy function $\tilde{H}(x, v) \equiv U(x) + \frac{1}{2} \langle v, v \rangle^{\mathcal{I}(\mathcal{M})}$ on $T\mathcal{I}(\mathcal{M})$, where $U \equiv V \circ \mathcal{I}^{-1} : \mathcal{I}(\mathcal{M}) \to \mathbb{R}$, since $H = \tilde{H} \circ \partial\mathcal{I}$. Hence we can reduce the problem of sampling from $\mathcal{M}$ to the problem of sampling from the embedded submanifold $\mathcal{I}(\mathcal{M})$ of $\mathcal{E}$. Thus from now on we consider the mechanics on the embedded submanifolds. In order to implement embedded HMC, the convenient reference measure on $\mathcal{I}(\mathcal{M})$ is chosen to be the Riemannian measure vol associated to the pullback metric $\langle \cdot, \cdot \rangle^{\mathcal{I}(\mathcal{M})}$. In other words our aim is

now to sample from $P \propto e^{-U}\text{vol}$.

The tangent space $T_{\iota(q)}\mathcal{E}$ above any point $q \in \mathcal{I}(\mathcal{M})$ decomposes into an orthogonal sum of the the horizontal tangent space $H_q \equiv \partial\iota(T_q\mathcal{I}(\mathcal{M}))$ and the vertical (or normal) tangent space consisting of $\langle\cdot,\cdot\rangle^{\mathcal{E}}$-orthogonal tangent vectors to H_q. We denote by $\text{hor} : T_{\iota(q)}\mathcal{E} \to H_q$ the horizontal projection for any $q \in \mathcal{I}(\mathcal{M})$. Samples from $\mathcal{N}\left(0, \langle\cdot,\cdot\rangle_q^{\mathcal{I}(\mathcal{M})}\right)$ may be obtained by sampling from $\mathcal{N}\left(0, \langle\cdot,\cdot\rangle_{\iota(q)}^{\mathcal{E}}\right)$ and taking the horizontal projection. Moreover we can express $\nabla^{\mathcal{I}(\mathcal{M})}U$ in terms of the embedding. To do so, we will need a differentiable extension $W : \mathcal{E} \to \mathbb{R}$, s.t. $\iota^*W = U$. Then

$$\partial\iota\nabla^{\mathcal{I}(\mathcal{M})}U = \text{hor}\big(\nabla^{\mathcal{E}}W\big)$$

which clearly does not depend on the choice of extension. In summary we have

$$\partial\mathcal{I}(\nabla^{\mathcal{M}}V) = \nabla^{\mathcal{I}(\mathcal{M})}U \qquad \partial\iota\nabla^{\mathcal{I}(\mathcal{M})}U = \text{hor}\big(\nabla^{\mathcal{E}}W\big).$$

Proof. For any $v \in T\mathcal{I}(\mathcal{M})$:

$$\iota^*\langle\nabla^{\mathcal{I}(\mathcal{M})}U, v\rangle = \langle\nabla^{\mathcal{I}(\mathcal{M})}(W\circ\iota), v\rangle^{\mathcal{I}(\mathcal{M})} = \partial\iota(v)(W) = \langle\nabla^{\mathcal{E}}W, \partial\iota(v)\rangle$$

and if $w \in \iota^*T\mathcal{E}$, with tangential component $\text{hor}(w) = \partial\iota(u)$ for some $u \in T\mathcal{I}(\mathcal{M})$, then

$$\begin{aligned}\langle\partial\iota(\nabla^{\mathcal{I}(\mathcal{M})}U), w\rangle &= \langle\partial\iota(\nabla_{\iota^*\langle,\rangle}U), \text{hor}(w)\rangle = \langle\partial\iota(\nabla^{\mathcal{I}(\mathcal{M})}U), \partial\iota(u)\rangle = \iota^*\langle\nabla^{\mathcal{I}(\mathcal{M})}U, u\rangle \\ &= \langle\nabla^{\mathcal{E}}W, \partial\iota(u)\rangle = \langle\text{hor}\big(\nabla^{\mathcal{E}}W\big), \partial\iota(u)\rangle = \langle\text{hor}\big(\nabla^{\mathcal{E}}W\big), w\rangle\end{aligned}$$

■

This calculations shows two things: the first one, is that in order to differentiate the potential U on $\mathcal{I}(\mathcal{M})$ in terms of the embedding, we need an extension of U defined on an open set of $\mathcal{E}$ (the result is independent of the extension) - it does not make sense to differentiate U directly in the embedding coordinates as was done in [66]. The second, is that if U arises from an extension defined on a neighbourhood in $\mathcal{E}$, then we never need to calculate the log-Jacobian terms present in [66], and this fact is unrelated to a lack of symplecticity as thought [160]. From the above theory we can implement HMC with target $P \propto e^{-V}\text{vol}_{\mathcal{M}}$ on $\mathcal{M}$ using for example the algorithm 7 based on the leapfrog integrator and full momentum refreshement. This generalises the algorithm of §2.1.2.

We now briefly discuss the special cases of hypersurfaces and the geodesics on the embedded manifolds in terms of the second fundamental form. We will then analyse some important special cases, starting with totally geodesic matrix manifolds §3.6.3, and then considering submanifolds of $\mathbb{R}^\ell$ in §3.6.4, and open submanifolds of vector spaces in §3.6.5.

Algorithm 7 HMC Algorithm to generate N samples $(x_r)_{i=1}^N$ from $\propto e^{-U}\mathrm{vol}_{\mathcal{I}(\mathcal{M})}$, and $(\mathcal{I}^{-1}(x_r))_{i=1}^N$ from $P \propto e^{-V}\mathrm{vol}_{\mathcal{M}}$.

procedure EMBEDDED-HMC(N)
  for for $r \in [1, N]$ **do**
    $q_0 \equiv x_{r-1}$
    sample $u_0 \sim \mathcal{N}\left(0, (\langle\cdot,\cdot\rangle_{q_0}^{\mathcal{E}})^{-1}\right)$
    $v_0 \leftarrow \mathrm{hor}(u_0)$
    $H_0 = V(q_0) + \frac{1}{2}\langle v_0, v_0\rangle_{q_0}^{\mathcal{E}}$
    $v_{\frac{1}{2}} = v_0 - \frac{1}{2}\delta t \mathrm{hor}\left(\nabla_{q_0}^{\mathcal{E}} W\right)$
    $(q_1, v_1^*) \equiv \Phi_{\delta t}^T(q_0, v_{\frac{1}{2}})$, i.e., follow the geodesic on $\mathcal{I}(\mathcal{M})$
    for for $m \in [1, k-1]$ **do**
      $v_{m+\frac{1}{2}} \equiv v_m^* - \delta t \mathrm{hor}\left(\nabla_{q_m}^{\mathcal{E}} W\right)$
      $(q_{m+1}, v_{m+1}^*) \equiv \Phi_{\delta t}^T(q_m, v_{m+\frac{1}{2}})$
    $v_k \equiv v_k^* - \frac{1}{2}\delta t \mathrm{hor}\left(\nabla_{q_k}^{\mathcal{E}} W\right)$
    $H_f = V(q_k) + \frac{1}{2}\langle v_k, v_k\rangle_{q_k}^{\mathcal{E}}$
    $\delta H \equiv H_f - H_0$
    $U \equiv \mathrm{Uniform}[0, 1]$
    if $U \leq \exp(-\delta H)$ **then** $x_r \equiv q_k$
    else $x_r \equiv q_0$
  return $(x_r)_{r=1}^N$

3.6.2.1 Second Fundamental Form, and Totally Geodesic Submanifolds

We now consider the geodesic problem on embedded manifolds. Let $\gamma : \mathbb{R} \to \mathcal{I}(\mathcal{M})$, with $\gamma(0) = q \in \mathcal{I}(\mathcal{M})$. The geodesic equation then reads

$$\frac{\nabla^{\mathcal{I}(\mathcal{M})}}{\mathrm{d}t}\dot{\gamma} = 0.$$

The image under the tangent map of the embedded tangent bundle $\partial\iota\Big(T\mathcal{I}(\mathcal{M})\Big)$ is the horizontal space (with respect to the ambient Riemannian metric) of the vector bundle $\iota^* T\mathcal{E}$ over $\mathcal{I}(\mathcal{M})$ and, under the identification ι, the connection on $T\mathcal{I}(\mathcal{M})$ equals the connection on that horizontal space (4.2 [89]). The connection on $\mathcal{I}(\mathcal{M})$ is then given by

$$\nabla^{\mathcal{I}(\mathcal{M})} = \nabla^{\mathcal{E}} - \mathcal{A}$$

where $\mathcal{A}$ is a "vector potential" known as the **second fundamental form**, with in particular $\mathcal{A}(\dot{\gamma}, \dot{\gamma}) = \mathcal{A}_\perp \nabla_{\dot{\gamma}}^{\mathcal{E}} \dot{\gamma}$, where $\mathcal{A}_\perp$ projects onto the vertical space to $\mathcal{I}(\mathcal{M})$. A curve γ on $\mathcal{E}$ will be

a geodesic of $\mathcal{I}(\mathcal{M}) \cong \iota\Big(\mathcal{I}(\mathcal{M})\Big)$ if (see also lemma 8.5 [208])

$$\frac{\nabla^{\mathcal{E}}}{\mathrm{d}t}\dot{\gamma} = \mathcal{A}(\dot{\gamma},\dot{\gamma}) \qquad \text{or equivalently} \qquad \mathrm{hor}(\frac{\nabla^{\mathcal{E}}}{\mathrm{d}t}\dot{\gamma}) = 0.$$

In particular if $\mathcal{E} = \mathbb{R}^n$, then the geodesic equation is simply $\mathrm{hor}(\ddot{\gamma}) = 0$, in other words the geodesics cannot accelerate in the tangential direction. This is because geodesics correspond to free motion, and thus any acceleration tangent to the manifold would correspond to a rate of change of potential energy. In contrast with the geodesics on the embedding space which do not accelerate at all, $\frac{\nabla^{\mathcal{E}}}{\mathrm{d}t}\dot{\gamma} = 0$, the geodesics of $\mathcal{I}(\mathcal{M})$ must accelerate in the normal direction in order to stay on the manifold. As in general relativity, we can view the trajectories of the geodesics on $\mathcal{I}(\mathcal{M})$ as the result of the curvature, or gauge field, $\mathcal{A}(\dot{\gamma},\dot{\gamma})$ which makes the free fall (motion with $V = 0$) on $\mathcal{I}(\mathcal{M})$ appear curved to an observer in $\mathcal{E}$. Submanifolds of $\mathcal{E}$ whose gauge field vanishes ($\mathcal{A} = 0$), so that the geodesics coincide with those of $\mathcal{E}$, are called **totally geodesic submanifolds** [285], and include the matrix groups discussed below.

3.6.2.2 Horizontal Projection on Hypersurfaces

When $\mathcal{I}(\mathcal{M})$ is a hypersurface of $\mathcal{E}$ with unit normal vector field n, we can use n to simplify the expressions of the second fundamental form. Indeed it holds that $\mathcal{A}_\perp X = X - \langle X, n\rangle^{\mathcal{E}} n$, and so $\mathcal{A}(X,Y) = \langle n, \nabla^{\mathcal{E}}_X Y\rangle n$ [208], and the tangential component for $\iota^* W = U$ is thus

$$\partial\iota \nabla^{\mathcal{I}(\mathcal{M})} U = \mathrm{hor}\big(\nabla^{\mathcal{E}} W\big) = \nabla^{\mathcal{E}} W - \big\langle \nabla^{\mathcal{E}} W, n\big\rangle^{\mathcal{E}} n.$$

Moreover if $\mathcal{I}(\mathcal{M}) = f^{-1}(c)$ is a level set of $f : \mathcal{E} \to \mathbb{R}$, then the unit normal vector field is $n = \frac{\nabla^{\mathcal{E}} f}{\|\nabla^{\mathcal{E}} f\|^{\mathcal{E}}}$; so $\mathrm{hor}\big(\nabla^{\mathcal{E}} W\big)$ in 7 may be replaced by

$$\mathrm{hor}\big(\nabla^{\mathcal{E}} W\big) = \nabla^{\mathcal{E}} W - \big\langle \nabla^{\mathcal{E}} W, \nabla^{\mathcal{E}} f\big\rangle^{\mathcal{E}} \frac{\nabla^{\mathcal{E}} f}{\langle \nabla^{\mathcal{E}} f, \nabla^{\mathcal{E}} f\rangle^{\mathcal{E}}}.$$

3.6.3 Matrix Groups

Consider a matrix subgroup $\iota : \mathcal{G} \to \mathrm{GL}(n)$ of $\mathrm{GL}(n)$, which is a closed submanifold. On $\mathrm{GL}(n)$ we use the left-invariant metric associated to the standard inner product $\langle X, Y\rangle_{\mathfrak{gl}(n)} \equiv \mathrm{tr}(X^\top Y)$ on the Lie algebra. Since $\mathrm{GL}(n)$ is an open subset of $\mathbb{R}^{n\times n}$ we can use the usual (matrix) global coordinates $x^{ij} : \mathrm{GL}(n) \to \mathbb{R}$, and identify as usual tangent vectors $M_{ij}\partial_{x^{ij}}|_x$ with matrices (M_{ij}), so $T\,\mathrm{GL}(n) \cong \mathrm{GL}(n) \times \mathfrak{gl}(n) \cong \mathrm{GL}(n) \times \mathbb{R}^{n\times n}$. The geodesic on $\mathrm{GL}(n)$ starting at (g, X)

is [253] (see §3.9.2.3 for a proof based on Euler–Arnold equations)

$$\gamma(t) = ge^{t(\Pi_{\mathcal{S}}X - \Pi_{\mathcal{A}}X)}e^{2t\Pi_{\mathcal{A}}X}, \tag{3.15}$$

where $\Pi_{\mathcal{S}}$ and $\Pi_{\mathcal{A}}$ are the symmetric and antisymmetric part of the matrix respectively. Let us now discuss the geodesic flow and vertical gradient step on some matrix subgroups.

Using ι we may identify $\mathfrak{g}$ with a subspace of $\mathfrak{gl}(n)$ and define the inner product by restriction $\langle \cdot, \cdot \rangle_{\mathfrak{g}} \equiv \iota^* \langle \cdot, \cdot \rangle_{\mathfrak{gl}(n)}$. For $\mathrm{SL}(n) \equiv \{g \in \mathrm{GL}(n) : \det g = 1\}$, it was shown in [265] that the geodesics of the induced metric were still given by (3.15). Specialising to $\mathrm{SO}(n)$, since $\mathfrak{so}(n)$ consists of antisymmetric matrices, the inner product becomes $\langle X, Y \rangle_{\mathfrak{so}(n)} = \mathrm{tr}(X^\top Y) = -\mathrm{tr}(XY)$, and by Ad-invariance the associated geodesic starting at (g, X) is given by ge^{tX}, again a special case of (3.15). It follows that both $\mathrm{SL}(n)$ and $\mathrm{SO}(n)$ are totally geodesic submanifolds of $\mathrm{GL}(n)$, so we can implement the geodesic motion in the ambient space without projection (see section 8.6 [337] for the complex matrix cases $\mathrm{U}(n), \mathrm{SU}(n)$ and $\mathrm{GL}(n, \mathbb{C})$).

From §3.9.1.3 we can show that given an extension $W : \mathrm{GL}(n) \to \mathbb{R}$,

$$g^{-1}\nabla_g^{\mathrm{GL}(n)} W = g^\top \nabla_x W, \qquad g^{-1}\nabla_g^{\mathcal{G}} W = \Pi_{\mathfrak{g}}[g^\top \nabla_x W],$$

where $\Pi_{\mathfrak{g}} : \mathfrak{gl}(n) \to \mathfrak{g}$ is the orthogonal projection (this follows directly by fixing an orthonormal bases of $\mathfrak{g}$ and applying the general formulae of §3.9.1.3). Hence we have $\mathrm{hor}(\nabla_g^{\mathrm{GL}(n)} W) = g\Pi_{\mathfrak{g}}[g^\top \nabla_x W] = \nabla_g^{\mathcal{G}} W$, since choosing any orthonormal basis (T_i) of $\mathfrak{g}$, it follows by left-invariance that (gT_i) is an orthonormal basis of $T_g\mathcal{G}$ and

$$\begin{aligned}\mathrm{hor}(\nabla_g^{\mathrm{GL}(n)} W) &= \left\langle \nabla_g^{\mathrm{GL}(n)} W, gT_i \right\rangle_g^{\mathrm{GL}(n)} gT_i = \left\langle g^{-1}\nabla_g^{\mathrm{GL}(n)} W, T_i \right\rangle_1^{\mathrm{GL}(n)} gT_i \\ &= \mathrm{tr}\left((g^\top \nabla_x W)^\top T_i\right) gT_i = g\Pi_{\mathfrak{g}}\left[g^\top \nabla_x W\right] = \nabla_g^{\mathcal{G}} W.\end{aligned}$$

This gives another interpretation of Lie group HMC, which was originally constructed using a symplectic structure on $\mathcal{G} \times \mathfrak{g}^*$ (see §3.3.1), and wherein we update the velocity on the Lie algebra by $\Pi_{\mathfrak{g}}[g^\top \nabla_x W]$, which is exactly the left trivialised version of the embedded HMC velocity update $\mathrm{hor}(\nabla_g^{\mathrm{GL}(n)} W)$.

Note that despite the fact that $\mathrm{GL}(n)$ is a curved manifold, using these updates we obtain an explicit HMC algorithm on standard matrix groups as a result of their totally geodesic embedding. This contrasts with the typical case of manifolds embedded in flat Euclidean space discussed next, where the geodesics are usually no longer "straight lines" in the ambient space, and may be difficult to compute.

3.6.4 HMC on Submanifolds of $\mathbb{R}^\ell$ with Geodesic Integrator

Let $\iota : \mathcal{I}(\mathcal{M}) \hookrightarrow \mathbb{R}^\ell$ be an embedded manifold, and suppose the target distribution has the form $P \propto e^{-U}\text{vol}$ where vol is the Riemannian volume form of $\iota^* \langle \cdot, \cdot \rangle^{\mathbb{R}^\ell}$. The potential energy U is assumed to be specified as the restriction of a function $W : \mathbb{R}^\ell \to \mathbb{R}$. Moreover for any vector $v \in T_x\mathbb{R}^\ell$ with $x \in \mathcal{I}(\mathcal{M})$, given an orthonormal basis $(n_i(x))$ of the normal space, the projection satisfies $(\mathcal{A}_\perp v)_i = v \cdot n_r (n_r)_i = (n_r)_i (n_r)_s v_s = N_{ir} N_{sr} v_s = (NN^\top v)_i$, i.e., $\mathcal{A}_\perp v = NN^\top v$ where $N_{ir} \equiv (n_r)_i$ is the matrix whose columns are the normal vectors. Hence the embedded scheme reduces to algorithm 8, and the Hausdorff density (and thus the determinant of Riemannian metric) never appears, contrasting with [66] (unlike what is stated in [160], this is true whether or not $NN^\top = \text{Id}$).

Algorithm 8 HMC Algorithm to generate N samples on $\mathcal{M} \subset \mathbb{R}^\ell$ from $P \propto e^{-V}\text{vol}_\mathcal{M}$, where $\text{vol}_\mathcal{M}$ is the surface measure induced by the inner product of $\mathbb{R}^\ell$, W is any differentiable extension of V defined in $\mathbb{R}^\ell$, and N is a matrix whose columns form an orthonormal basis of the normal space (see (3.16) for the case $\mathcal{M} = f^{-1}(0)$).

procedure EMBEDDED-HMC(N)
  for for $r \in [1, N]$ **do**
    $q_0 \equiv x_{r-1}$
    sample $u_0 \sim \mathcal{N}(0, I_{\ell\times\ell})$
    $v_0 \leftarrow \left(I_{\ell\times\ell} - N(q_0)N(q_0)^\top\right) u_0$
    $H_0 = V(q_0) + \frac{1}{2} v_0^\top v_0$
    $v_{\frac{1}{2}} = v_0 - \frac{1}{2}\delta t \left(I_{\ell\times\ell} - N(q_0)N(q_0)^\top\right) \nabla^{\mathbb{R}^\ell}_{q_0} W$
    $(q_1, v_1^*) \equiv \Phi^T_{\delta t}(q_0, v_{\frac{1}{2}})$, i.e., follow the geodesic on $\mathcal{I}(\mathcal{M})$
    for for $m \in [1, k-1]$ **do**
      $v_{m+\frac{1}{2}} \equiv v_m^* - \delta t \left(I_{\ell\times\ell} - N(q_m)N(q_m)^\top\right) \nabla^{\mathbb{R}^\ell}_{q_m} W$
      $(q_{m+1}, v_{m+1}^*) \equiv \Phi^T_{\delta t}(q_m, v_{m+\frac{1}{2}})$
    $v_k \equiv v_k^* - \frac{1}{2}\delta t \left(I_{\ell\times\ell} - N(q_k)N(q_k)^\top\right) \nabla^{\mathbb{R}^\ell}_{q_k} W$
    $H_f = V(q_k) + \frac{1}{2} v_k^\top v_k$
    $\delta H \equiv H_f - H_0$
    $U \equiv \text{Uniform}[0, 1]$
    if $U \leq \exp(-\delta H)$ **then** $x_r \equiv q_k$
    else $x_r \equiv q_0$
  return $(x_r)_{r=1}^N$

In particular if $\mathcal{I}(\mathcal{M}) = f^{-1}(0)$ for some function $f : \mathbb{R}^\ell \to \mathbb{R}^{\ell-n}$, then $\left(\nabla^{\mathbb{R}^\ell} f^1, \ldots, \nabla^{\mathbb{R}^\ell} f^{\ell-n}\right)$ span the normal space, and (recall $(\partial f)_{ij} = \partial_j f^i$) then

$$NN^\top = \partial f^\top \left(\partial f \partial f^\top\right)^{-1} \partial f. \tag{3.16}$$

The example of the Stiefel manifold is discussed in [66]. Observe however in this case we

are using the Riemannian metric on Stiefel manifold obtained by restricting the Euclidean metric, which is non-canonical and does not respect the natural symmetries of the manifold (a correction term is necessary if the target distribution is expressed in terms of the uniform measure). Closed-form expressions for the corresponding non-canonical geodesics may be found in 2.2.2 [108]. On the other hand, the algorithm obtained by symplectic reduction in §3.8 is based on the uniform measure.

3.6.5 Open Submanifolds of Vector Spaces

If $\iota : \mathcal{M} \to V$ is an open submanifold of a finite-dimensional vector space V, then $\partial\iota\Big(T_q\mathcal{M}\Big) = V$, and we may identify $T\mathcal{M} \cong \mathcal{M} \times V$, that is the horizontal tangent space at any point is the whole embedding tangent space V. Under this identification the Riemannian metric on V and its restriction to $\mathcal{M}$ are essentially identical, and the horizontal projection reduces to the identity map. Since a vector space is a manifold with global unconstrained coordinates (once we have chosen a basis), we can use these to differentiate the extension of functions on $\mathcal{M}$. If we further have closed-form expressions for the geodesics then HMC may be easily implemented.

As an example we consider the implementation proposed in [161], and applied to learning the spectral density matrix for vector time series, for the space of positive definite symmetric matrices $\mathbf{SPD}(n)$, since [161] states that, in contrast with [66], their algorithm is intrinsic and does not rely on an isometric embedding. $\mathbf{SPD}(n)$ is an open submanifold of $S(n)$, the vector space of symmetric matrices, so $T_q\mathbf{SPD}(n) \cong S(n)$. Potential energies on $\mathbf{SPD}(n)$ are usually given by the restriction of an extended potential W on $S(n)$. If x^i are global coordinates on $S(n)$ which select (in some order) the independent components of the matrix, we denote by $\upsilon : S(n) \to \mathbb{R}^{\frac{1}{2}n(n+1)}$ the map $\upsilon(S) \equiv \big(x^1(S), \dots, x^{\frac{1}{2}n(n+1)}(S)\big)$. Let $\mathrm{vec} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n^2}$ be the map which given a matrix outputs the column vector obtained by stacking its columns. There exists a "duplication" matrix $D_n \in \mathbb{R}^{n^2\times\frac{1}{2}n(n+1)}$ s.t., for any $S \in S(n)$

$$\mathrm{vec}(S) = D_n\upsilon(S).$$

The expressions for the canonical metric and its gradient used in [161] are precisely their extrinsic form in these coordinates: the canonical Riemannian metric of $\mathbf{SPD}(n)$ in these coordinates are (using (5) [268])

$$\langle A, B\rangle_P^{\mathbf{SPD}(n)} \equiv \upsilon(A)^\top G(P)\upsilon(B), \qquad \text{where} \qquad G(P) \equiv D_n^\top\,(P\otimes P)\,D_n,$$

while the Riemannian gradient in the embedding coordinates is (see (30) [268])

$$\nabla_P^{\mathbf{SPD}(n)} W = D_n^+ (P \otimes P) D_n^{+\top} \partial_x W$$

where $D_n^+ \in \mathbb{R}^{\frac{1}{2}n(n+1)\times n^2}$ is the Moore-Penrose inverse, and $\partial_x W$ is the column vector $(\partial_{x^i} W)$. Since we also have closed-form expressions for the geodesics in the canonical metric, see §3.1.3.1.1, we can then implement HMC on $\mathbf{SPD}(n)$ using for example the geodesic integrator in §3.5.1. Another example is the open Stiefel manifold of k linearly independent vectors in $\mathbb{R}^n$ (rather than orthonormal k-frames), which can be viewed as an open submanifold of $\mathbb{R}^{n\times k}$, and thus may be equipped with the Riemannian metric induced from $\mathbb{R}^{n\times k}$.

3.7 HMC via Riemannian Submersions

In §3.6 we have discussed the use of embeddings $\mathcal{I} : \mathcal{M} \to \mathcal{E}$ to implement HMC on $\mathcal{M}$. We have seen that to obtain the Riemannian gradient on the manifold we needed to compute the Riemannian gradient of an extension in the embedding and project it to the horizontal tangent space. Similarly, unless the embedding is totally geodesic, we saw that the geodesics on $\mathcal{M}$ differed from those in $\mathcal{E}$ due to the action of the second fundamental form. We now briefly discuss the implementation of HMC on $\mathcal{M}$ using submersions $\tau : \mathcal{E} \to \mathcal{M}$. We will show that these are more natural maps to implement HMC, since they preserve the Riemannian gradients and geodesics. Submersions typically arise on shape spaces $\mathcal{M} \equiv \mathcal{E}/\mathcal{K}$ associated to group actions, and we shall discuss this case in detail in §3.8.

Let $\tau : \mathcal{E} \to \mathcal{M}$ be a surjective Riemannian submersion between complete manifolds, that is τ and $\partial_g \tau$ are surjective for each $g \in \mathcal{E}$. We define the **vertical space** $V_g\mathcal{E} \equiv \mathrm{ver}_g \equiv \ker \partial_g \tau$, and **horizontal space** $H_g\mathcal{E} \equiv \mathrm{hor}_g \equiv (V_g\mathcal{E})^\perp$, which gives a decomposition of the tangent bundle $T\mathcal{E} = H(\mathcal{E}) \oplus V(\mathcal{E})$. Moreover, the restriction of the tangent map to the horizontal spaces $\partial_g \tau : H_g(\mathcal{E}) \to T_{\tau(g)}\mathcal{M}$ is an isometry. We saw in §3.6.1 that isometries are the isomorphisms of Hamiltonian mechanics in velocity space (they are symplectomorphism), and thus preserve the mechanics. The difference now is that we only have an isometry pointwise (on each tangent space) when we restrict to the horizontal spaces. Indeed τ is not required to be a diffeomorphism and will usually not be injective. Let us now discuss the relation between the geodesic motions and Riemannian gradients.

The geodesics of $\mathcal{M}$ are related to those in $\mathcal{E}$ by the following result: a horizontal curve $\gamma : \mathbb{R} \to \mathcal{E}$ is a geodesic of $\mathcal{E}$ iff $\tau \circ \gamma$ is a geodesic of $\mathcal{M}$, and every geodesic of $\mathcal{M}$ can be characterised this way by Lemma 26.11 [264]. Hence we may obtain the geodesics of $\mathcal{M}$ by

simply projecting the horizontal geodesics of $\mathcal{E}$. Moreover by Lemma 26.12, every geodesic with initial velocity in $H(\mathcal{E})$ will be horizontal. This is in clear contrast with the Riemannian embeddings case, in which there was no clear way to obtain the geodesics on $\mathcal{M}$ from the geodesics of its embedding, and a geodesic of $\mathcal{E}$ with initial velocity in the horizontal space would stay on $\mathcal{M}$ iff the embedding was totally geodesic. Denote by V the potential energy on $\mathcal{M}$, i.e., we aim to sample from the target distribution $P \propto e^{-V}\text{vol}$ where vol is the Riemannian measure on $\mathcal{M}$. We define its extension to be the pullback $W \equiv \tau^* V : \mathcal{E} \to \mathbb{R}$. The Riemannian gradients of V and W are preserved under the submersion

$$\partial\tau(\nabla^{\mathcal{E}}_g W) = \nabla^{\mathcal{M}}_{\tau(g)} V.$$

and in fact $\nabla^{\mathcal{E}}_g W \in H(\mathcal{E})$, the Riemannian gradient is automatically horizontal:

Proof. First note that if γ is a vertical curve, then $\left\langle \nabla^{\mathcal{E}}_g W, \dot{\gamma} \right\rangle^{\mathcal{E}} = \mathrm{d}W(\dot{\gamma}) = \partial V \circ \partial\tau(\dot{\gamma}) = \partial V \circ 0 = 0$. Thus $\nabla^{\mathcal{E}}_g W \in H(\mathcal{E})$. Moreover, for any $u_{\tau(g)} \in T_{\tau(g)}\mathcal{M}$, there exists $w_g \in T_g\mathcal{E}$ s.t. $\partial_g\tau(w_g) = u_{\tau(g)}$. Therefore, by the definition of the pull-back metric

$$\begin{aligned}\left\langle \partial_g\tau\big(\nabla^{\mathcal{E}}_g W\big), u_{\tau(g)} \right\rangle^{\mathcal{M}} &= \left\langle \partial_g\tau\nabla^{\mathcal{E}}_g(V\circ\tau), \partial_g\tau(w_g) \right\rangle^{\mathcal{M}} = \left\langle \nabla^{\mathcal{E}}_g(V\circ\tau), w_g \right\rangle^{\mathcal{E}} = \mathrm{d}_g(V\circ\tau)(w_g) \\ &= \mathrm{d}_{\tau(g)}V \circ \partial_g\tau(w_g) = \mathrm{d}_{\tau(g)}V(u_{\tau(g)}) = \left\langle \nabla^{\mathcal{M}}_{\tau(g)} V, u_{\tau(g)} \right\rangle^{\mathcal{M}}.\end{aligned}$$

■

It follows that we can implement geodesic integrators on $\mathcal{M}$ by implementing them on $\mathcal{E}$, as long as we use the extension W as our potential energy and we restrict the heat bath to the horizontal spaces, that is:

- sampling velocity on $H_g\mathcal{E}$ is equivalent to sampling the velocity on $T_{\tau(g)}\mathcal{M}$;
- given $v \in H_g\mathcal{E}$ the potential flow $\Phi^W_{\delta t}$ maps it to $v - \delta t \nabla^{\mathcal{E}}_g W$, whose projection corresponds to updating $\partial_g\tau(v)$ according to $\Phi^V_{\delta t}$;
- the geodesic flow on $T\mathcal{E}$ with initial conditions (g, v), projects to the geodesic flow on $T\mathcal{M}$ starting at $(\tau(g), \partial_g\tau(v))$.

In §3.8 we will discuss in detail the important case where the submersions arise from a Lie group action, that is $\mathcal{M}$ is a shape space. In that case the algorithm is a special case of the general theory of reduction by symmetry, where the horizontal space is determined by the mechanical connection and momentum map. We finish this short section with two famous examples of Riemannian submersions.

3.7.1 $\tau : \mathbf{S}(n,k) \to \mathbf{G}(n,k)$

We can define an obvious submersion from the Stiefel manifold $\mathbf{S}(n,k)$ of k-frames in $\mathbb{R}^n$ to the Grassmannian manifold $\mathbf{G}(n,k)$ of k-dimensional vector spaces in $\mathbb{R}^n$ by $\tau\big(f_1,\dots,f_k\big) \equiv \operatorname{span}\big(f_1,\dots,f_k\big)$, which is Riemannian when these spaces are equipped with their canonical metric [4]. Thus we can obtain an HMC scheme on Grassmannians from an HMC scheme of Stiefel manifolds. In fact many optimisation algorithms on Grassmannian manifolds are expressed in terms of the "identification" $\tau(g) \in \mathbf{G}(n,k)$ with a representative element $g \in \tau^{-1}(\tau(g))$. We may think of [162] as a hybrid scheme which uses this identification together with the geodesics on $\mathbf{G}(n,k)$, which can be obtained by projecting the geodesics of $\mathbf{S}(n,k)$.

3.7.2 Hopf Fibration $\tau : S^{2n+1} \to \mathbb{CP}^n$

Consider $\mathcal{E} = S^{2n+1}$ with the canonical metric from $\mathbb{R}^{2n+2} \cong \mathbb{C}^{n+1}$. The complex projective space is defined as $\mathbb{CP}^n \equiv \{[z] : z \in S^{2n+1}\}$ where the equivalence class $[z]$ is defined by identifying elements of $S^{2n+1} = \{z \in \mathbb{C}^{n+1} : \|z\|^2 = |z_1|^2 + \cdots + |z_{n+1}|^2 = 1\}$ related by rotation $e^{i\theta} \in S^1 \cong U(1)$, so $z \sim w$ iff $z = e^{i\theta}w \equiv \big(e^{i\theta}w_1,\dots,e^{i\theta}w_{n+1}\big)$ for some $\theta \in \mathbb{R}$. The projection $\tau : S^{2n+1} \to \mathbb{CP}^n$, $z \mapsto [z]$, is called the Hopf map. By requiring it to be a Riemannian submersion, it induces the canonical Riemannian metric on $\mathbb{CP}^n$, called the Fubini-Study metric (this allows us to obtain expressions for the geodesics of $\mathbb{CP}^n$ [126]).

3.8 HMC on Shape Spaces by Reduced Dynamics

In §3.7 we explained how to express on $\mathcal{E}$ the steps implementing a HMC sampler on $\mathcal{M}$ with a geodesic integrator using the Riemannian submersion. An important supply of submersions are induced by the action of a Lie group, and in that case we can apply results from geometric mechanics to implement the sampler. Specifically, given a free and proper action of $\mathcal{K}$ on $\mathcal{E}$, the projection onto the quotient $\tau : \mathcal{E} \to \mathcal{E}/\mathcal{K}$ is a submersion, and the quotient manifold $\mathcal{M} = \mathcal{E}/\mathcal{K}$ is known as a **shape space**. Moreover if we choose a metric on $\mathcal{E}$ which is $\mathcal{K}$-invariant, it defines a canonical metric on $\mathcal{E}/\mathcal{K}$ for which τ is a Riemannian submersion. We can then apply the algorithm of §3.7 to sample on $\mathcal{E}/\mathcal{K}$. In this section we will consider the continuous mechanics, since splitting methods are compositions of continuous mechanics. We will see that the mechanics on $\mathcal{E}/\mathcal{K}$ can be obtained from the mechanics on $\mathcal{E}$ using Noether's theorem (or symplectic reduction), and preserves the form of the Hamiltonians, in the sense

that it naturally relates the Riemannian Newton systems adapted to the target measure and its lift. Moreover, understanding mechanics with symmetries is important to derive integrators that may be efficiently implemented.[14]

In §3.8.1 we will describe the general procedure of symplectic reduction in order to describe the mechanics on $\mathcal{M}$ in terms of mechanics with $\mathcal{K}$-symmetries on $\mathcal{E}$. To do so we will introduce the geometric mechanics description of the horizontal and vertical spaces on $\mathcal{E}$ in terms of the mechanical connection, and momentum map. Such a reduction requires setting the "conserved momentum" to 0, though in §3.8.1.3 we will see that magnetic HMC arises by reduction of a mechanics that conserves a non-vanishing quantity, which suggests using the electromagnetic symplectic structure in future work to improve the ergodicity of HMC (interestingly, the momentum flip is not an antisymplectic involution in magnetic HMC, and one must flip the magnetic field). We mention that in physics, symplectic reduction is typically used to study Hamiltonians with symmetries, by first reducing the mechanics to the reduced symplectic space (a bundle over the quotient space with an "electromagnetic" symplectic structure), and then reconstructing the original system from the reduced one. Instead, in HMC we start from an arbitrary Hamiltonian associated to the log-density, and construct a Hamiltonian with symmetries on an appropriate space over which the implementation of the algorithm is simpler. This method may be useful also for the systems with holonomic constraints, common in molecular dynamics (see §3.5.2), using the impetus-striction formalism [356], combined with explicit integrators for non-separable Hamiltonian [335]. Here we will instead focus on geodesic orbit manifolds, which include hyperspheres, hyperbolic spaces, Stiefel manifolds, Grassmann manifolds, $\mathrm{SU}(n)$, $\mathrm{SO}(n)$, the spaces of positive definite matrices, as well as their complex and hyperbolic versions, which are common in statistics and lattice QCD, and allows us to improve and extend the methods proposed in [162, 161, 204, 66, 186]. Thus, in sections §3.8.2 and §3.8.3 we will specialise to the case of a homogeneous manifolds, and obtain explicit algorithms on naturally reductive manifolds, which essentially reduce HMC to the problem of efficiently implementing the Lie group exponential. Moreover, since the resulting scheme is entirely on a Lie group, we can use their special properties, such as the fact the symplectic structure and Hamiltonian flows have global coordinate-independent expressions (see §3.9), to tune the integrator as in lattice QCD.

[14] **The principle of mediocrity**: professor José M. Figueroa-O'Farrill once told me that the cosmological principle, which states that "at sufficiently large scales, the Universe looks the same to all observers", or mathematically that the universe is a homogeneous manifold, should really be thought as a statement about the mediocrity of the universe. Luckily, this mediocrity is often the reason why we can make calculations about the universe. Similarly, in HMC the tractable numerical integrators are typically those that are composed of invariant (or homogeneous) vector fields. In this light, we observe that the reason why RMHMC [131] is difficult to implement is that it replaces a bi-invariant metric with a metric with no symmetries. Several authors rely on the distinction between separable and non-separable Hamiltonians to explain this problem, but it is easy to define separable Hamiltonians on Lie groups with intractable flows.

3.8.1 Noether's Theorem and Reduced Dynamics

When a Lie group $\mathcal{K}$ (left/right) acts freely and properly on a manifold $\mathcal{E}$, with action denoted Φ_k (or sometimes "$k\cdot$"), the projection $\tau : \mathcal{E} \to \mathcal{E}/\mathcal{K}$ is a submersion [2, proposition 4.1.23].[15] Thus each Φ_k is a bijection of $\mathcal{E}$, and we assume it is smooth, $\Phi_k \in \mathrm{Diff}(\mathcal{E})$. It then follows that the lifted action $(\Phi_k)^*$ acts symplectically on $T^*\mathcal{E}$. The horizontal and vertical spaces can be described in terms of the infinitesimal actions (or action fields) $\xi_{\mathcal{E}} \in \Gamma(T\mathcal{E})$ generated by elements $\xi \in \mathfrak{k}$ of the Lie algebra of $\mathcal{K}$, defined by[16]

$$\xi_{\mathcal{E}}(q) \equiv \frac{\mathrm{d}}{\mathrm{d}t} e^{t\xi} \cdot q\big|_{t=0}, \tag{3.17}$$

where $e^{t\xi}$ is the Lie group exponential. These are the vectors that are mapped to 0 under the quotient projection $\mathcal{E} \to \mathcal{M}$, as can be seen from the fact that $\gamma : t \mapsto e^{t\xi} \cdot q$ is a curve that lies entirely in the coset $\mathcal{K} \cdot q$, and thus $\tau \circ \gamma$ is a constant curve on $\mathcal{M}$. Thus the vertical space is

$$\mathrm{ver}_q = \{\xi_{\mathcal{E}}(q) : \xi \in \mathfrak{k}\}.$$

In order to describe the horizontal space, we introduce the **momentum map** of the action, $J : T^*\mathcal{E} \to \mathfrak{k}^*$, where

$$\langle J(p_q), \xi \rangle_* = \langle p_q, \xi_{\mathcal{E}}(q) \rangle_* , \qquad \forall \xi \in \mathfrak{k}$$

where $\langle \cdot, \cdot \rangle_*$ denotes the dual pairing. From section 4.2 [2] we see that J generalises the usual notions of momentum. Indeed if $\mathcal{K} = \mathcal{E} = \mathbb{R}^n$ acts on itself by translation, $k \cdot q = k + q$, then for any $\xi \in \mathfrak{k} = \mathbb{R}^n$, $\xi_{\mathcal{E}}(q) = \xi$ and J is the usual linear momentum $J(q, p) = p$. On the other hand if $\mathcal{K} = \mathrm{SO}(3, \mathbb{R})$ acts on $\mathcal{E} = \mathbb{R}^3$ by matrix multiplication, then we recover the angular momentum $J(q, v) = q \times v$. More generally J describes the conserved quantities/currents induced by the symmetries of the Hamiltonian system as we now describe by introducing the geometric generalisation of the standard Noether's theorem.

3.8.1.1 Noether's Theorem

In the Hamiltonian perspective, **Noether's theorem** states that if $H : T^*\mathcal{E} \to \mathbb{R}$ is invariant under the (lifted) action of $\mathcal{K}$, $H(k \cdot (q, p)) = H(q, p)$, then J is conserved by the flow of X_H (theorem 2.7.2 [250], or 4.2.2 [2]). In terms of Lagrangians $\mathcal{L} : T\mathcal{E} \to \mathbb{R}$, and momentum map $J^{\mathcal{L}} : T\mathcal{E} \to \mathfrak{k}^*$ defined by $\langle J(v_q), \xi \rangle_* = \langle \mathbb{F}\mathcal{L}(v_q), \xi_{\mathcal{E}}(q) \rangle_*$, Noether's theorem states that

[15] Note that we need the action of $\mathcal{K}$ on $\mathcal{E}$ to be free and proper to ensure that the quotient $\mathcal{E}/\mathcal{K}$ is a manifold. For example this is the case when $\mathcal{K}$ is a closed subgroup of a Lie group $\mathcal{E}$.

[16] I.e., $\xi_{\mathcal{E}}(q) = \partial_1 \Phi(q)[\xi]$

for the Euler-Lagrange system associated to a $\mathcal{K}$-invariant Lagrangian $J^{\mathcal{L}}$ is constant in time (independently of whether $\mathcal{L}$ is regular) [250]. The usual Noether's theorem is a special case of the above result, and may be recovered as follows. Consider the action of $\mathcal{K} \equiv \mathbb{R}$ on $\mathcal{E} \equiv \mathbb{R}^n$ by translation of the i^{th} coordinate, $k \cdot (q^1, \dots, q^n) \equiv (q^1, \dots, k + q^i, \dots, q^n)$. Then $\xi_{\mathbb{R}^n}(q) = (0, \dots, \xi, \dots, 0)$ and $J(q, v) = \frac{\partial \mathcal{L}}{\partial v^i}$. Thus, if $\mathcal{L}$ does not depend on the i^{th} coordinate (called a cyclic or ignorable coordinate), then its corresponding conjugate momentum defined as $p_i \equiv \frac{\partial \mathcal{L}}{\partial v^i}$ is conserved along a solution of Euler-Lagrange equation. This can also be seen directly from Euler-Lagrange equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial \mathcal{L}}{\partial v^i}\right)\Big|_{\dot{q}} = \frac{\partial \mathcal{L}}{\partial q^i} = 0, \qquad \Rightarrow \qquad \frac{\partial \mathcal{L}}{\partial \dot{q}^i} = \text{constant}.$$

In general if we expand $\xi_{\mathcal{E}}(q) \equiv (\xi^a X_a)_{\mathcal{E}}(q) = \xi^a (X_a)_{\mathcal{E}}(q) \equiv \xi^a K_a^i(q) \partial_{q^i}$ using bases (∂_{q^i}) and (X_a) of $\mathfrak{k}$ and $T_q\mathcal{E}$ respectively, then the momentum map J becomes

$$J(q, v) = \frac{\partial \mathcal{L}}{\partial v^i} K_a^i(q) X^a$$

which is called **Noether's current** [250].

In fact as we discuss next, not only is J conserved by X_H, but appropriate quotients of the level sets of J define a reduced symplectic manifold, on which the projection of X_H defines a reduced Hamiltonian system. Importantly for the context of HMC, when we set $J = 0$, then this reduced symplectic manifold is precisely $T^*\mathcal{M}$ (with canonical symplectic structure), and with an appropriate choice of H the reduced Hamiltonian corresponds to our target system, so that the latter may then be implemented using X_H.

3.8.1.2 Reduced Mechanics on $\mathcal{E}/\mathcal{K}$

While the vertical space is uniquely defined by the submersion, the horizontal space depends on a choice of connection on the (principal $\mathcal{K}$-) bundle $\tau : \mathcal{E} \to \mathcal{E}/\mathcal{K}$. As mentioned earlier, an appropriate choice of Riemannian metric on $\mathcal{E}$ is one for which $\mathcal{K}$ acts by isometries, since this will define a $\mathcal{K}$-invariant kinetic energy on $T\mathcal{E}$, and any such metric defines a quotient metric $\langle \cdot, \cdot \rangle^{\mathcal{M}}$ on $\mathcal{M} = \mathcal{E}/\mathcal{K}$ for which τ is a Riemannian submersion [125, theorem 22.14]. In physics the connection associated to this Riemannian metric on $\mathcal{E}$, i.e., to the horizontal space described in §3.7, is called the **mechanical connection** $\mathcal{A}$. In terms of the momentum map we have

$$\mathrm{hor}_q = \{(q, v) : J((q, v)^\flat) = 0\}.$$

Indeed we note that $(q,v) \in \ker(J \circ \flat)$ iff $0 = \left\langle (q,v)^\flat, \xi_{\mathcal{E}}(q) \right\rangle_* = \langle (q,v), \xi_{\mathcal{E}}(q) \rangle_{\mathcal{E}}$, that is (q,v) is orthogonal to the vertical subspace. The connection 1-form $\mathcal{A} : T\mathcal{E} \to \mathfrak{k}$ associated to this choice of horizontal space is characterised by

$$\mathcal{A}(q,v) = \xi \text{ if } (q,v) = \xi_{\mathcal{E}}(q), \qquad \mathcal{A}(q,v) = 0 \text{ if } (q,v) \in \mathrm{hor}_q,$$

and extending by linearity. From this we immediately see that we have

$$\mathrm{ver}_q(q,v) = \big(\mathcal{A}(q,v)\big)_{\mathcal{E}}(q),$$

so that we may use $\mathcal{A}$ to project to the horizontal space

$$\mathrm{hor}_q(q,v) = (q,v) - \big(\mathcal{A}(q,v)\big)_{\mathcal{E}}(q).$$

Now given a target P on $\mathcal{M}$, consider the Hamiltonian on $T\mathcal{M}$ adapted to P (recall §3.3.2), with kinetic energy induced by the quotient metric. Given the potential energy V on $\mathcal{M}$ (corresponding to the log-density of P), the extended potential $W \equiv \tau^* V$ on $\mathcal{E}$ is $\mathcal{K}$-invariant.[17] Hence, the resulting Hamiltonian $W + \frac{1}{2}\|\cdot\|_{\mathcal{E}}^2$ on $\mathcal{E}$ is then invariant under the induced action of $\mathcal{K}$ on $T^*\mathcal{E}$ (indeed both the kinetic and potential on $\mathcal{E}$ are $\mathcal{K}$-invariant). It follows by Noether's theorem that J is constant along the flow of this $\mathcal{K}$-invariant Hamiltonian. In other words, if we start at an initial point (q,v) that satisfies $J((q,v)^\flat) = \mu \in \mathfrak{k}^*$, then the flow will remain on the level set $J^{-1}(\mu)$. In particular, if $\mu = 0 \in \mathfrak{k}^*$ is a regular value of J (which means $\partial_z J$ is surjective for any $z \in J^{-1}(0)$), then the motion will remain on the horizontal space $J^{-1}(0)$, the gauge $\mathcal{A}$ will act trivially, $\mathcal{A} = 0$, and the reduced Hamiltonian system will evolve on a reduced symplectic manifold that is symplectomorhpic to $T^*\mathcal{M}$ [286, 2].[18] Indeed, we have

$$J^{-1}(0)/\mathcal{K} \cong T^*\mathcal{M} \cong T\mathcal{M},$$

and since $\mathcal{A} = 0$, the Hamiltonian dynamics on $\mathcal{E}$ will reduce to a Hamiltonian dynamics on $\mathcal{M}$ induced by a reduced Hamiltonian $V + \frac{1}{2}\langle \cdot, \cdot \rangle^{\mathcal{M}}$ [250, proposition 3.4.1], which is adapted to the target measure $P \propto e^{-V}\mathrm{vol}_{\mathcal{M}}$ as desired. Hence we can generate the mechanical trajectories of $V + \frac{1}{2}\langle \cdot, \cdot \rangle^{\mathcal{M}}$ entirely from mechanical trajectories in $\mathcal{E}$. See Fig 3.3 for an illustration.

[17] Note the property $\mathcal{L}_{\xi_{\mathcal{E}}} W = 0$ for any $\xi_{\mathcal{E}}$, since

$$\mathcal{L}_{\xi_{\mathcal{E}}} W(q) = \frac{\mathrm{d}}{\mathrm{d}t} W(e^{t\xi} \cdot q)\big|_0 = \frac{\mathrm{d}}{\mathrm{d}t} V \circ \pi(e^{t\xi} \cdot q)\big|_0 = \frac{\mathrm{d}}{\mathrm{d}t} V \circ \mathcal{K} \cdot (e^{t\xi} \cdot q)\big|_0 = \frac{\mathrm{d}}{\mathrm{d}t} V\big(\mathcal{K} \cdot q\big)\big|_0 = 0.$$

This yields another proof of the fact the momentum update preserves "horizontality", since $\langle \nabla W, \xi_{\mathcal{E}} \rangle^{\mathcal{E}} = \mathrm{d}W(\xi_{\mathcal{E}}) = \mathcal{L}_{\xi_{\mathcal{E}}} W = 0$.

[18] Observe that the lift of a free and proper action is free and proper [308], so all the conditions hold to apply the symplectic reduction theorem

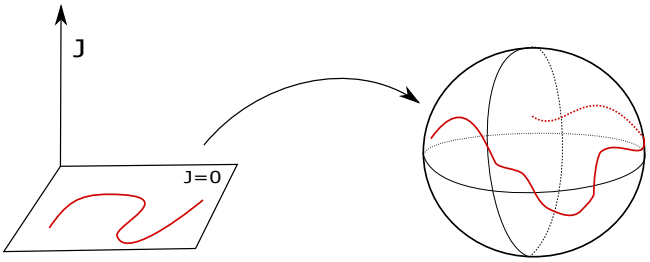

Figure 3.3: Example of reduced trajectory for a mechanical system with SO(2) symmetry on $T^*\,\mathrm{SO}(3)$; by symmetry, if initially $J = 0$ then the system remains on the level set $J = 0$ and the reduced trajectory is a Hamiltonian dynamics on the sphere $\mathrm{SO}(3)/\,\mathrm{SO}(2)$.

The theory of symplectic reduction allows for far more general systems than the one discussed above, and for example in [28] we have discussed how to implement mechanical trajectories on the Lorentzian single sheeted hyperbolic space (though it is not possible to use the Lorentzian metric to refresh momentum for the purpose of HMC). We also mention that the reason why the Deep MCMC method proposed recently in [317] is not volume-preserving stems precisely from ignoring the theory of symplectic reduction, that provides conditions under which a Hamiltonian system with symmetries reduces to a Hamiltonian system on the quotient.

To summarise, we have shown that symplectic reduction can be used to implement the MD step when the Riemannian metric on $\mathcal{E}$ is chosen to be $\mathcal{K}$-invariant. We now consider in §3.8.2 the question of what other conditions we can impose on the metric in order to make the geodesic equations as simple as possible. We will see that this question has a natural answer in the case of a reductive homogeneous space, but we briefly explain in §3.8.1.3 how to recover magnetic HMC [342] by reducing an appropriate standard mechanical system with $\mu \neq 0$. This also shows that while the standard magnetic HMC cannot be obtained from canonical mechanics by some choice of Hamiltonian (lemma 5 [342]), it may be obtained by symplectic reduction of a canonical system.

3.8.1.3 Electromagnetic Symplectic Structure and Magnetic HMC

Reduction of mechanics can also be used to implement non-canonical mechanics on the reduced space when the conserved quantity does not vanish $\mu \neq 0$. In that case the reduced symplectic structure gains an additional magnetic two-form, and the reduced Hamiltonian must be shifted via the mechanical connection. We will focus on the case $\mathcal{M} = \mathbb{R}^n$, which we will view as the shape space of $\mathcal{E} = \mathbb{R}^n \times S^1$. Specifically we consider the action of $\mathcal{K} \equiv S^1 \cong U(1)$ on $\mathcal{E}$,

$$e^{i\theta} \cdot (x, e^{i\vartheta}) \equiv (x, e^{i(\vartheta+\theta)}).$$

Since $S^1 \cong U(1)$ the Lie algebra of S^1 is $\mathfrak{u}(1) = i\mathbb{R}$. The action fields follow immediately, $(i\theta)_{\mathcal{E}}(x, e^{i\vartheta}) = \frac{\mathrm{d}}{\mathrm{d}t}(x, e^{i(\vartheta+t\theta)})|_{t=0} = (0, i\theta e^{i\vartheta}) \cong (0, i\theta)$ where in the last step we have identified $T_\varphi S^1 = T_{e^{i\varphi}} U(1) \cong \mathfrak{u}(1)$ by left-trivialisation (so $T_{(q,\varphi)}\mathcal{E} = \mathbb{R}^n \times \mathfrak{u}(1)$). Further identifying $\mathfrak{u}(1) \cong \mathbb{R}$, we have $T^*\mathcal{E} = T^*\mathbb{R}^n \oplus T^*S^1 \cong \mathbb{R}^n \times \mathbb{R}^n \times S^1 \times \mathbb{R} \cong \mathbb{R}^n \times S^1 \times \mathbb{R}^n \times \mathbb{R}$, and the momentum map $J : \mathbb{R}^n \times S^1 \times \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ is simply $J(x, e^{i\theta}, p, \alpha) = \alpha$. We define the **magnetic potential** [19]

$$\mathcal{A} \equiv A_i \mathrm{d}x^i + \mathrm{d}\theta : T\mathcal{E} \cong \mathbb{R}^n \times S^1 \times \mathbb{R}^n \times \mathbb{R} \to \mathfrak{k} \cong \mathbb{R}$$

and the **magnetic form**

$$F = \mathrm{d}\mathcal{A} = \mathrm{d}A$$

which we view as a 2-form on $\mathbb{R}^n$. Then the symplectic reduction theorem implies that the canonical phase space $T^*\mathcal{E}$ (with canonical symplectic structure) reduces for any $\mu \in \mathbb{R}$ to the symplectic manifold $J^{-1}(\mu)/S^1 \cong T^*\mathbb{R}^n$ with symplectic structure $\omega_A \equiv \mathrm{d}x \wedge \mathrm{d}p - \mu F$ [286]. The symplectic structure ω_A is often called the **electromagnetic symplectic structure** since F and A generalise the **electromagnetic field** and its **vector potential** from electromagnetic theory. Indeed, consider the two-form $F \equiv B + \mathrm{d}t \wedge E$ on $\mathbb{R}^{3+1}$, where B and E are a two-form and a 1-form on $\mathbb{R}^3$ (the magnetic and electric fields). Define $A \equiv (\mathbf{A}, -\phi)$, where ϕ is the $\mathbb{R}$-valued electric potential and $\mathbf{A}$ the $\mathbb{R}^3$-valued magnetic potential, then $F = \mathrm{d}A$ yields the usual equations relating the fields B, E to their potentials,

$$\mathbf{E} = -\nabla\phi - \frac{\partial \mathbf{A}}{\partial \mathbf{t}} \qquad \mathbf{B} = \nabla \times \mathbf{A},$$

while $\mathrm{d}F = \mathrm{d}\mathrm{d}A = 0$ yields Maxwell's equations (see §A.5.7.1)

$$\nabla \cdot \mathbf{B} = \mathbf{0} \qquad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial \mathbf{t}}.$$

From the derivations in §A.5.7.1, we see that Hamilton's equations $i_{\hat{H}}\omega_A = \mathrm{d}H$ with the electromagnetic symplectic structure now implies

$$\dot{z} = \begin{pmatrix} 0 & I \\ -I & K \end{pmatrix} \partial_z H = \begin{pmatrix} \partial_p H \\ -\partial_q H + K\partial_p H \end{pmatrix},$$

for an antisymmetric matrix K representing the magnetic field (see also exercise 2.4-1 [251]). These are precisely the equations of motion used in an HMC algorithm called **magnetic HMC** and proposed in [342], where it was shown that in some cases the presence of the magnetic field can improve mixing of the HMC chain. Indeed as shown in B.2 [342] the flows Φ^V and Φ^T are

[19] This is a connection 1-form since $\mathcal{A}(\theta_{\mathcal{E}}) = \mathcal{A}(\theta\partial_\theta) = \theta$ and obeys the equivariant condition.

still tractable (for $T(p) \equiv \frac{1}{2}p^\top p$), and thus may be combined to form an appropriate numerical integrator. However some additional care must be taken to handle the fact the usual momentum flip is no longer an antisymplectic involution for ω_A (recall §3.4.1).

3.8.2 Homogeneous Manifolds

Submersions induced by an action typically arise when $\mathcal{M}$ is a homogeneous $\mathcal{G}$- manifold, i.e., there is a Lie group $\mathcal{E} \equiv \mathcal{G}$ that (left) acts transitively on $\mathcal{M}$, and we now specialise to this setting. To realise $\mathcal{M}$ as a quotient space, we fix an origin $p_0 \in \mathcal{M}$, and let $\mathcal{K} \equiv \mathcal{K}_{p_0}$ denote its isotropy group. Then $\mathcal{M} \cong \mathcal{G}/\mathcal{K}$, where we identify $g \cdot p_0 \sim g\mathcal{K}$, and $g\mathcal{K} = \mathcal{K} \cdot g$ denotes the left coset induced by the right action $k \mapsto R_k$ of $\mathcal{K}$ on $\mathcal{G}$, which is free and proper by proposition 22.4 [125]. The projection $\tau : \mathcal{G} \to \mathcal{M}$, $g \mapsto g\mathcal{K}$ is then a submersion, which we make Riemannian as explained in §3.8.1 by defining a (right) $\mathcal{K}$-invariant metric on $\mathcal{G}$. Then 0 is a regular value of the associated momentum map J by proposition 2.9.1 [250], and

$$\xi_{\mathcal{G}}(g) = \frac{\mathrm{d}}{\mathrm{d}t} e^{t\xi} \cdot g|_{t=0} = \frac{\mathrm{d}}{\mathrm{d}t} g e^{t\xi}|_{t=0} = \partial_1 L_g(\xi),$$

so the action fields are simply the left-invariant vector fields induced by $\mathfrak{k}$. The metric identifies $T^*\mathcal{G} \cong T\mathcal{G}$ as usual. It is convenient to restrict to metrics generated by $\mathrm{Ad}_{\mathcal{G}}\mathcal{K}$-invariant inner products on $\mathfrak{g}$ (here $\mathrm{Ad}_{\mathcal{G}}\mathcal{K} \equiv \mathrm{Ad}\,|_{\mathcal{K}}$), so that we can work in left-trivialised (or body) coordinates, that is we identify $T\mathcal{G} \cong \mathcal{G} \times \mathfrak{g}$, via $(g, v_g) \mapsto (g, \theta_g v_g)$, where $\theta \in \Gamma(\mathcal{G}, T^*\mathcal{G} \otimes \mathfrak{g})$ is the **Maurer-Cartan form** defined by $\theta : g \mapsto \partial_g L_{g^{-1}}$ [2]. In these coordinates the induced action of $\mathcal{K}$ on $\mathcal{G} \times \mathfrak{g}$ becomes $\partial R_k(g, \xi) = (gk, \mathrm{Ad}_{k^{-1}}\,\xi)$. [20] Let us now discuss the implementation of HMC for naturally reductive manifolds in these body coordinates.

3.8.3 Naturally Reductive Homogeneous Spaces

Suppose $\mathcal{G}/\mathcal{K}$ is reductive with reductive decomposition $\mathfrak{g} = \mathfrak{p} \oplus \mathfrak{k}$, and assume we have a $\mathcal{G}$-invariant metric on $\mathcal{M}$, which we extend to an $\mathrm{Ad}_{\mathcal{G}}\mathcal{K}$-invariant inner product on $\mathfrak{g}$ for which $\mathfrak{p} \perp \mathfrak{k}$ (recall §3.1.2). This defines a left $\mathcal{G}$-invariant and right $\mathcal{K}$-invariant Riemannian metric on $\mathcal{G}$, and τ is a Riemannian submersion. The associated horizontal space at g is simply $\partial_1 L_g \mathfrak{p} \subset T_g\mathcal{G}$ (since isometries preserve orthogonality), and in body coordinates the momentum

[20] Denoting by $\sim$ the identification, we have

$$\begin{aligned}\partial R_k(g, \xi) \sim \partial R_k\big(g, \partial_1 L_g \xi\big) &= \big(gk, \partial_g R_k \circ \partial_1 L_g \xi\big) = \big(gk, \partial_1 (R_k \circ L_{gk} \circ L_{k^{-1}})\xi\big) \\ &= \big(gk, \partial_1 L_{gk} \circ \partial_1 (R_k \circ L_{k^{-1}})\xi\big) = \big(gk, \partial_1 L_{gk} \circ \mathrm{Ad}_{k^{-1}}\,\xi\big) \sim (gk, \mathrm{Ad}_{k^{-1}}\,\xi).\end{aligned}$$

map simplifies to[21]

$$J : \mathcal{G} \times \mathfrak{g} \to \mathfrak{k}, \quad J(g, \xi) = \xi_{\mathfrak{k}},$$

so the horizontal bundle is thus $J^{-1}(0) \cong \mathcal{G} \times \mathfrak{p}$ in body coordinates.[22] Moreover noting that if $(g, X) \in J^{-1}(0)$, then $\partial\tau \circ \partial R_k(g, X) = \partial\tau \circ (gk, \mathrm{Ad}_{k^{-1}} X) = (\mathrm{Ad}_{gk} \circ \mathrm{Ad}_{k^{-1}} X)_{\mathcal{M}}(gk\mathcal{K}) = (\mathrm{Ad}_g X)_{\mathcal{M}}(g\mathcal{K})$, which is independent of $k \in \mathcal{K}$, we obtain the isomorphism between the symplectic manifolds of interest

$$J^{-1}(0)/\mathcal{K} \cong T\mathcal{M}, \quad \text{with} \quad [(g, X)] \mapsto \partial_g\tau \circ \partial_1 L_g(X).$$

Suppose now that $\mathcal{G}/\mathcal{K}$ is naturally reductive. We calculate the flows Φ^W and Φ^T from which we can compute a multi-stage numerical integrator §2.1.3. Assume for the moment that $\mathcal{G}$ is a Lie group of real matrices (we consider the more general case in §A.5.7.2), with kinetic energy T defined by the inner product $(X, Y) \mapsto \frac{1}{2}\operatorname{tr}\left(X^\top Y\right)$ on $\mathfrak{g} \subset \mathfrak{gl}(n) = \mathbb{R}^{n\times n}$. The map $\Phi^T_t : \mathcal{G} \times \mathfrak{p} \to \mathcal{G} \times \mathfrak{p}$ is (see §3.9.2.2)

$$\Phi^T_t(g, X) \equiv \left(ge^{tX}, X\right).$$

Note in particular that the Lie algebra-valued velocity X remains fixed during this update (in contrast to the reduced dynamics). If (T_i) is an orthonormal basis of $\mathfrak{g}$, let $G_{ij} \equiv \frac{1}{2}\operatorname{tr}(T_i^\top T_j) = \delta_{ij}$, and suppose the lifted potential energy $\tau^* V$ on $\mathcal{G}$ is obtained by restriction of a function $W : \mathrm{GL}(n) \to \mathbb{R}$, we have (see §3.9.1.2)

$$\begin{aligned}
\widehat{W} &= -e_i|_g(W)T^i = -\mathrm{d}W(gT_i)\frac{T_i}{G_{ii}} = -\frac{\partial W}{\partial x_{ab}}\mathrm{d}x_{ab}(gT_i)T_i \\
&= -\operatorname{tr}\left(\nabla_x W^\top gT_i\right)T_i = -\operatorname{tr}\left((g^\top\nabla_x W)^\top T_i\right)T_i \\
&= -2\operatorname{tr}\left((g^\top\nabla_x W)^\top \frac{T_i}{\sqrt{2}}\right)\frac{T_i}{\sqrt{2}} = -2\Pi(g^\top\nabla_x W)
\end{aligned}$$

where Π is the orthogonal projection onto $\operatorname{span}(T_i)$ with respect to the Frobenius inner product, and $x_{ab} : \mathrm{GL}(n, \mathbb{R}) \to \mathbb{R}$ are the canonical matrix coordinates. Hence

$$\Phi^W_t(g, X) = \left(g, X - 2t\Pi(g^\top\nabla_x W)\right),$$

[21] since for any $\nu \in \mathfrak{k}$, $\langle J(g, \xi), \nu\rangle \equiv \langle J(g, \partial_1 L_g \xi), \nu\rangle_* = \langle \partial_1 L_g \xi, \nu_{\mathcal{G}}(g)\rangle^{\mathcal{G}}_g = \langle \partial_1 L_g \xi, \partial_1 L_g \nu\rangle^{\mathcal{G}}_g = \langle \xi, \nu\rangle^{\mathcal{G}}_1 = \langle \xi_{\mathfrak{k}}, \nu\rangle^{\mathcal{K}}_1$

[22] The mechanical connection $\mathcal{A} : \mathcal{G} \times \mathfrak{g} \to \mathfrak{k}$ is also particularly simple in body coordinates, and is closely related to the Maurer-Cartan form,

$$\mathcal{A}(g, \xi) = \xi \text{ iff } \xi \in \mathfrak{k}, \quad \text{and} \quad \mathcal{A}(g, \xi) = 0 \text{ iff } \xi \in \mathfrak{p}.$$

Sample Space $\mathcal{M}$	$\mathcal{M} \equiv \mathcal{G}/\mathcal{K}, \quad \tau : g \mapsto g\mathcal{K}$	$\mathfrak{p}$
Hypersphere	$\mathrm{SO}(n)/\mathrm{SO}(n-1)$ $g \mapsto (g_{i1})_i$	$\begin{pmatrix} 0 & -v^\top \\ v & 0_{(n-1)\times(n-1)} \end{pmatrix}$ $v \in \mathbb{R}^{n-1}$
Stiefel	$\mathrm{SO}(n)/\mathrm{SO}(n-k)$ $g \mapsto (g_{ir})_{i,r\in[1,k]}$	$\begin{pmatrix} R & -M^\top \\ M & 0 \end{pmatrix}$ $R \in \mathfrak{so}(k), M \in \mathbb{R}^{(n-k)\times k}$
Complex Stiefel	$\mathrm{U}(n)/\mathrm{U}(n-k)$ $g \mapsto (g_{ir})_{i,r\in[1,k]}$	$\begin{pmatrix} R & -M^\dagger \\ M & 0 \end{pmatrix}$ $R \in \mathfrak{u}(k), M \in \mathbb{C}^{(n-k)\times k}$
Grassmannian	$\mathrm{SO}(n)/(\mathrm{SO}(k)\times\mathrm{SO}(n-k))$ $g \mapsto \mathrm{span}\left((g_{ir})_{i,r\in[1,k]}\right)$	$\begin{pmatrix} 0 & -M^\top \\ M & 0 \end{pmatrix}$ $M \in \mathbb{R}^{(n-k)\times k}$
Symmetric PD	$\mathrm{GL}(n)^+/\mathrm{SO}(n)$ $g \mapsto gg^\top$	$n \times n$ symmetric matrices
Hermitian PD	$\mathrm{GL}(\mathbb{C}^n)/\mathrm{U}(n)$ $g \mapsto gg^\dagger$	$n \times n$ Hermitian matrices

Table 3.2: Some examples of naturally reductive manifolds, and their decomposition.

so we obtain the HMC algorithm 9. In particular when $\Pi(A) = \Pi_{\mathcal{A}}(A) \equiv \frac{1}{2}(A - A^\top)$, then $\Pi_{\mathcal{A}}(A) = -\Pi_{\mathcal{A}}(A^\top)$ so $\hat{V} = 2\Pi_{\mathcal{A}}(\partial_x V g)$. Similarly if $\Pi(A) \equiv \Pi_{\mathcal{S}}(A) = \frac{1}{2}(A + A^\top)$, then $\Pi_{\mathcal{S}}(A) = \Pi_{\mathcal{S}}(A^\top)$ and $\hat{V} = -2\Pi_{\mathcal{S}}(\partial_x V g)$. This may be applied to many important manifolds in statistics, such as hyperspheres, Stiefel manifolds, Grassmann manifolds, and the space of covariance matrices (as well as their complex and hyperbolic versions), see table 3.2. Of course it is not always easy to visualise the algorithm in those spaces, but we provide below a visual example on the sphere, and additional experiments on hyperbolic spaces may be found in [28].

3.8.3.1 Example: Sphere

Let us briefly describe the parametrisation of the sphere as a homogeneous space. The standard action (by matrix product) of $\mathrm{SO}(3,\mathbb{R})$ on $S_2 \subset \mathbb{R}^3$ is transitive, and the isotropy group of the "north pole" $p_0 = (1,0,0)$ is $\mathcal{K} \equiv \mathcal{K}_{p_0} \equiv \{\mathrm{diag}(1, A) \text{ for } A \in \mathrm{SO}(2,\mathbb{R})\}$. Any coset $g\mathcal{K}_{p_0}$ may be identified with the point $g \cdot p_0 \in \mathcal{M}$. The sphere is then diffeomorphic to $\mathrm{SO}(3,\mathbb{R})/\mathrm{SO}(2,\mathbb{R})$,

Algorithm 9 HMC Algorithm to generate N samples, given $q_0 \in \mathcal{G}$, with leapfrog integrator of k leapfrog trajectories of time step δt.

procedure HomMC(N)
  for for $r \in [1, N]$ **do**
    $g_0 \equiv q_{r-1}$
    sample a matrix v_0 in $\mathfrak{p}$ with $\mathcal{N}(0, I_{n\times n})$
    $H_0 \equiv W(g_0) + \frac{1}{4}\operatorname{tr}(v_0^\top v_0)$
    $v_{\frac{1}{2}} \equiv v_0 - \delta t \Pi(g_0^\top \nabla_x W)$
    $g_1 \equiv g_0 e^{\delta t v_{1/2}}$
    for for $m \in [1, k-1]$ **do**
      $v_{m+\frac{1}{2}} \equiv v_{m-\frac{1}{2}} - 2\delta t \Pi(g_m^\top \nabla_x W)$
      $g_{m+1} \equiv g_m e^{\delta t v_{m+1/2}}$
    $v_k \equiv v_{k-\frac{1}{2}} - \delta t \Pi(g_k^\top \nabla_x W)$
    $H_f \equiv W(g_k) + \frac{1}{4}\operatorname{tr}(v_k^\top v_k)$
    $\delta H \equiv H_f - H_0$
    $U \equiv \text{Uniform}[0, 1]$
    if $U \leq \exp(-\delta H)$ **then** $q_r \equiv g_k$
    else $q_r \equiv g_0$
  return $(\tau(q_r))_{r=1}^N$

or, in other words, any point on the sphere can be identified with a rotation about a line in the 2–3 plane. We choose the generators (basis of the Lie algebra) of $\mathfrak{so}(3)$ to be $(T_i)_{jk} = \varepsilon_{ijk}$, i.e.,

$$T_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}, \; T_2 = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \text{and} \quad T_3 = \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

In this case $\mathfrak{k} = \operatorname{span} T_1$ generates $\mathrm{SO}(2, \mathbb{R})$ and $\mathfrak{p} = \operatorname{span}(T_2, T_3)$ generates $\exp(\mathfrak{p}) \cdot p_0 = \mathcal{M}$. In order to clarify the geometrical meaning of these generators we parameterise the coset space $\mathcal{M}$ in terms of real angles θ and ϕ,

$$S(\theta, \phi) \equiv \exp\left[\theta(\sin\phi\, T_2 - \cos\phi\, T_3)\right] = \exp\left[\theta \begin{pmatrix} 0 & -\cos\phi & -\sin\phi \\ \cos\phi & 0 & 0 \\ \sin\phi & 0 & 0 \end{pmatrix}\right]$$

$$= \begin{pmatrix} \cos\theta & -\sin\theta\cos\phi & -\sin\theta\sin\phi \\ \sin\theta\cos\phi & (\sin\phi)^2 + \cos\theta(\cos\phi)^2 & (\cos\theta - 1)\cos\phi\sin\phi \\ \sin\theta\sin\phi & (\cos\theta - 1)\cos\phi\sin\phi & (\cos\phi)^2 + \cos\theta(\sin\phi)^2 \end{pmatrix},$$

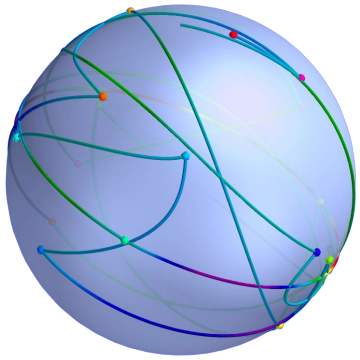

Figure 3.4: 20 HMC trajectories, chronologically coloured blobs, the trajectories are coloured proportionally to the values of δH, with potential $V(x, y, z) = yz^2 \exp(x^2)$, $\delta t = 0.1$, $\tau = 2$.

and these coordinates allows us to identify the point

$$S(\theta, \phi) \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \cos\theta \\ \sin\theta\cos\phi \\ \sin\theta\sin\phi \end{pmatrix} \in S_2$$

with the spherical polar coordinate (θ, ϕ). Similarly we may parameterise the isotropy group $\mathcal{K}_{p_0}$ as

$$\exp(tT_1) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos t & \sin t \\ 0 & -\sin t & \cos t \end{pmatrix};$$

clearly, $\exp(tT_1)p_0 = p_0$. Note we have constructed a parametrisation $S(\theta, \phi)\exp(tT_1)$ of $\mathrm{SO}(3, \mathbb{R})$, while $S(\theta, \phi)\exp(tT_1)S(\theta, \phi)^T$ provides a parametrisation of the isotropy group $\mathcal{K}_p$ at $p \equiv S(\theta, \phi)p_0$.

Of course in practice we never use such coordinates, instead we implement directly HMC on the Lie group SO(3) and then project the samples to the sphere. In Fig. 3.4 we generate some trajectories on the sphere for $V(x, y, z) \equiv yz^2 \exp(x^2)$ using mechanics on SO(3) and symplectic reduction, while in Fig. 3.5 we sample the world.

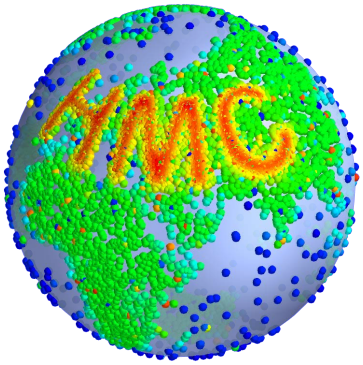

Figure 3.5: *HMC via reduction by symmetry for a well-known distribution*

3.8.3.2 Complex Matrices

Consider now the case where $\mathcal{G}$ is a real Lie group of complex matrices, with inner product on $\mathfrak{g}$ obtained by restricting the standard inner product $X, Y \mapsto \frac{1}{2}\Re \operatorname{tr}\left(X^\dagger Y\right)$ on $\mathbb{C}^{n\times n}$.[23] Recall we assume the potential is provided as a map $W : \mathrm{GL}(\mathbb{C}^n) \to \mathbb{R}$, and use the fact $\mathrm{GL}(\mathbb{C}^n)$ is an open submanifold of the (real $2n^2$-dimensional) manifold $\mathbb{C}^{n\times n}$ with coordinates z_{ab}. Tangent vectors to $\mathrm{GL}(\mathbb{C}^n)$ are identified with matrices as in the real case: $M^{ab}\partial_{z_{ab}} \sim (M^{ab})$, and in particular $\mathrm{d}z_{cd}(M^{ab}\partial_{z_{ab}}) = M^{cd}$. Recall also that $\mathrm{d}W \equiv \frac{\partial W}{\partial z_{ab}}\mathrm{d}z_{ab} + \frac{\partial W}{\partial \bar{z}_{ab}}\mathrm{d}\bar{z}_{ab} = \frac{\partial W}{\partial z_{ab}}\mathrm{d}z_{ab} + \overline{(\frac{\partial \bar{W}}{\partial z_{ab}})}\mathrm{d}\bar{z}_{ab} = \frac{\partial W}{\partial z_{ab}}\mathrm{d}z_{ab} + \overline{(\frac{\partial W}{\partial z_{ab}})}\mathrm{d}\bar{z}_{ab}$. Moreover from $\mathrm{d}z = \mathrm{d}x + i\mathrm{d}y$ and $\mathrm{d}\bar{z} = \mathrm{d}x - i\mathrm{d}y$ we have $\mathrm{d}\bar{z}(v) = \overline{\mathrm{d}z(v)}$. Denote by (T_i) an orthonormal basis of generators and Π the projection onto $\mathfrak{g}$. In particular for the Lie algebra of $\mathrm{U}(n)$ of anti-hermitian matrices $\Pi_{\mathfrak{u}(n)}(A) = \frac{1}{2}(A - A^\dagger)$, while for the Lie algebra of $\mathrm{SU}(n)$ of traceless Hermitian matrices $\Pi_{\mathfrak{su}(n)}(A) = \frac{1}{2}(A - A^\dagger) - \operatorname{tr} A/n$. In order to implement the HMC momentum update in this case we need to replace the one obtained in §3.8.3 with the following:

$$\begin{aligned}
\widehat{W} &= -\mathrm{d}W(e_i)T^i = -\mathrm{d}W(e_i)T_i = -\frac{\partial W}{\partial z_{ab}}\mathrm{d}z_{ab}(gT_i)T_i - \overline{(\frac{\partial W}{\partial z_{ab}})}\mathrm{d}\bar{z}_{ab}(gT_i)T_i \\
&= -\nabla_{z_{ab}} W(gT_i)_{ab}T_i - \overline{\nabla_{z_{ab}} W(gT_i)}_{ab}T_i = -2\Re \operatorname{tr}\left(\partial_z W gT_i\right)T_i = -2\Re \operatorname{tr}\left((g^\dagger \partial_z W^\dagger)^\dagger T_i\right)T_i \\
&= -4\Re \operatorname{tr}\left((g^\dagger \partial_z W^\dagger)^\dagger \frac{T_i}{\sqrt{2}}\right)\frac{T_i}{\sqrt{2}} = -4\Pi(g^\dagger \partial_z W^\dagger).
\end{aligned}$$

In the appendix §A.5.7.2 we discuss the general case in which we implement the equations of motion on a naturally reductive manifold $\mathcal{G}$ using a faithful representation $\psi : \mathcal{G} \to \mathrm{GL}(n)$.

[23] Given a complex vector space with Hermitian inner product $\langle\cdot,\cdot\rangle$ becomes a real inner product space by restricting the field to $\mathbb{R}$ and using the inner product $\Re\langle\cdot,\cdot\rangle$.

3.8.3.3 *Invariance of the Shadow Hamiltonian*

Finally, let us briefly explain why splitting integrators retain the symmetries of the Hamiltonians that compose them. Since the action of $\mathcal{K}$ on $\mathcal{G} \times \mathfrak{g}$ is symplectic it is also a Poisson map [166, proposition 4.33], which means $\{f \circ \Phi_k, g \circ \Phi_k\}_{\mathcal{G}\times\mathfrak{g}} = \{f, g\}_{\mathcal{G}\times\mathfrak{g}} \circ \Phi_k$. Hence the Poisson bracket of $\mathcal{K}$-invariant functions is a $\mathcal{K}$-invariant function. If we use a symplectic integrator defined by splitting the Hamiltonian $H = H_1 + \cdots + H_r$, the corresponding shadow Hamiltonian will be a sum of Poisson brackets of the Hamiltonians H_i. It follows that if each Hamitlonian H_i is $\mathcal{K}$-invariant, then so is the shadow Hamiltonian. Hence integrators that are built by alternating between the flows of W and T are automatically horizontal (the trajectory stays in $\mathcal{G} \times \mathfrak{p}$) since both W and T are $\mathcal{K}$-invariant. This is still true for the force-gradient integrators. Indeed, from the $\mathrm{ad}_\mathfrak{k}$-invariance of the inner product on $\mathfrak{g}$, we have $c_{rij} = -c_{jir}$ for any $i \in \mathfrak{k}$ (the structure constants always satisfy $c^i_{jk} = -c^i_{kj}$). Moreover (if W, T are invariant under the right action of $\mathcal{K}$ on $\mathcal{G}$) we have $\{W, \widehat{\{W, T\}}\} = -2G^{jk}G^{ls}e_l(W)e_j e_s(W)\xi_k \in \mathfrak{p}$ (see §3.9 for notation and §3.9.3.4 for proof).

3.9 Hamiltonian Monte Carlo On Lie Groups

HMC was originally constructed for computations in lattice QCD, and while the equations of motion were initially informally obtained, a rigorous derivation was given soon after [186] in order to formally construct HMC for lattice gauge systems on compact semisimple Lie groups. Poisson brackets, shadow Hamiltonians and higher order integrators based on these equations were further discussed in [187]. The extension was designed by constructing a symplectic structure (and thus a mechanics) on $\mathcal{G} \times \mathfrak{g}^*$, but the relation with the standard mechanics induced by the Liouville form on $T^*\mathcal{G}$ was never explained - indeed the authors were unaware of the relation with the Euler–Poincaré formalism we discuss below.

In this section we formalise the HMC schemes given in [187, 186] and connect them to standard mechanics with symmetries on Lie groups. In the context of QCD, the position update relies on matrix exponentiation [187, 87, 186, 192] which is derived by solving the equations of motion on semisimple lie algebras, assuming that the Killing form induces a positive definite inner product on the Lie algebra, which is not the case in general and as a result cannot be used for HMC (due to the momentum refreshment step). Indeed, this holds for compact semisimple Lie groups (for example $\mathrm{SO}(n)$ and $\mathrm{SU}(n)$) but is not the case for most other Lie groups, such as $\mathrm{GL}(n)$ or $\mathrm{SL}(n)$. In fact the matrix exponentiation simply corresponds to the geodesic flow

of the Killing form in left-trivialised coordinates, and the connection with geometric mechanics allows us to remove the compactness assumption on the Lie algebra (or equivalently, the assumption that the Killing form is negative definite). In §3.9.1.1 we show the symplectic structure used in [187, 186] is symplectomorphic to the canonical one on $T^*\mathcal{G}$. It follows that many results from geometric mechanics that assume standard mechanics, such as symplectic reduction on cotangent bundles, can be used in HMC [2]. Indeed the HMC algorithm via symplectic reduction we introduced previously relies on the fact that the cotangent-lift of diffeomorphism on $\mathcal{G}$ is a symplectomorphism with respect to the canonical symplectic structure, so that $\mathcal{K}$ acts symplectically on $T^*\mathcal{G}$.
Since HMC is usually implemented on the tangent bundle, we derive the corresponding symplectic structure on $\mathcal{G} \times \mathfrak{g}$ induced by a metric. As mentioned above, in [187, 186] the kinetic flow was simply given by left-invariant vector fields as a result of the bi-invariance of the Killing-form. This corresponds to solving the reconstruction equation for a constant curve on $\mathfrak{g}$, but, as we explain in §3.9.1.2, the geodesic flow of more general left-invariant metrics must solve both the Euler-Arnold and reconstruction equations. In §3.9.1.3 we provide the explicit relation between the leapfrog scheme on Lie groups as used in [187] and the general one on manifolds. In §3.9.2 we then consider the Euler-Arnold equation in more detail, and explain how it can be simplified and solved by choosing inner products with appropriate symmetries. In particular we derive the associated geodesic motion on $\mathrm{GL}(n)$, semi-simple Lie groups, as well as the geodesic formula that we used to implement HMC on naturally reductive manifolds.

3.9.1 Hamiltonian Monte Carlo on Lie Groups

In section §3.9.1.1 we explain the global construction of symplectic structures on Lie groups, and derive the symplectic structures on $\mathcal{G} \times \mathfrak{g}^*$, $T\mathcal{G}$ and $\mathcal{G} \times \mathfrak{g}$ induced by the canonical 1-form on $T^*\mathcal{G}$. Using the latter symplectic structure we will derive the global equations for the Hamiltonian vector fields §3.9.1.2 , and discuss the Euler–Poincaré formalism. Finally we shall discuss geodesic integrators §3.9.1.3. Let us fix a basis (ξ_i) of $\mathfrak{g}$, denote by $e_i \in \mathfrak{g}_L$ the induced left-invariant vector fields (so $e_i(1) = \xi_i$), by θ^i the dual 1-forms, $\theta^i(e_j) = \delta^i_j$, and by θ the Maurer-Cartan form.

3.9.1.1 Symplectic Structures on Lie Groups

As with any symplectic manifold we can find local Darboux coordinates (x^i, p_i) for $T^*\mathcal{G}$ in which the Liouville form reads $\eta = p_i \mathrm{d}x^i$. The induced symplectic structure is then $\omega = -\mathrm{d}\eta =$

$\mathrm{d}x^i \wedge \mathrm{d}p_i$, and Hamilton's equations take the familiar form

$$\dot{x} = \frac{\partial H}{\partial p}, \qquad \dot{p} = -\frac{\partial H}{\partial x}.$$

However on manifolds there are in general no global Darboux coordinates. Fortunately Lie groups have the useful property that we can instead construct η in a coordinate-independent way using the non-vanishing 1-forms θ^i of the Maurer–Cartan frame. Following [11] we define the momentum as $J \equiv \left(\theta^{-1}\right)^* : T^*\mathcal{G} \to \mathfrak{g}^*$. Here the inverse is taken pointwise, so setting $J|_g \equiv J|_{T^*_g\mathcal{G}}$ and $\theta_g^{-1} \equiv (\theta_g)^{-1} : \mathfrak{g} \to T_g\mathcal{G}$, we have $J|_g = (\theta_g^{-1})^*$. Thus for any $\alpha_g \in T^*_g\mathcal{G}$,

$$J|_g(\alpha_g) = (\theta_g^{-1})^*(\alpha_g) = \alpha_g \circ \theta_g^{-1} = \alpha_g \circ \partial_1 L_g = L_g^* \alpha_g \in \mathfrak{g}^*,$$

where we have used $\theta_g^{-1} = \partial_1 L_g$ (due to the inverse function theorem). It follows that J is the momentum map associated to the right action of $\mathcal{G}$ on itself (see §3.8.1). We can "expand" θ and J using the basis (ξ_i), which yields $\theta = \xi_i \otimes \theta^i$ and $J = \theta^i(1) \otimes p_i$ where we have defined the momentum coordinates $p_i \equiv \langle J(\cdot), \xi_i \rangle_* : T^*\mathcal{G} \to \mathbb{R}$. We then define the Liouville 1-form by $\eta \equiv p_i \pi^* \theta^i \in \Gamma(T^*\mathcal{G}, T^*T^*\mathcal{G})$ on $T^*\mathcal{G}$, and its associated symplectic structure is its exterior derivative[24]

$$\omega \equiv -\mathrm{d}\eta = -\mathrm{d}(p_i \pi^* \theta^i) = \underbrace{\pi^*\theta^i \wedge \mathrm{d}p_i}_{\text{usual "}\mathrm{d}x \wedge \mathrm{d}p\text{" term}} + \underbrace{\tfrac{1}{2} p_i c^i_{jk} \pi^*\theta^j \wedge \pi^*\theta^k}_{\text{additional non-abelian term}} \tag{3.18}$$

On $\mathcal{G} = \mathbb{R}^n$ by the Maurer-Cartan equation and Poincaré lemma we have $\theta^i = \mathrm{d}x^i$, so that η may be viewed as a generalisation of the usual Liouville form $p_i \mathrm{d}x^i$. The symplectic structure above was the one constructed in [186] to define a mechanics on group manifolds in order to sample distributions on SU(3) (for lattice QCD). [25] In fact $\omega \equiv -\mathrm{d}\eta$ is the canonical 2-form since η has the following property which shows η is the canonical Liouville form: for any 1-form $\beta : \mathcal{G} \to T^*\mathcal{G}$

$$\beta^* \eta = \beta. \tag{3.19}$$

In other words, the mechanics introduced in [186] is equivalent to canonical mechanics.

Proof. Let $\beta \in \Gamma(T^*\mathcal{G})$, $p_i : T^*\mathcal{G} \to \mathbb{R}$, $\pi : T^*\mathcal{G} \to \mathcal{G}$ and $\eta \in \Lambda^1(T^*\mathcal{G})$. By definition of the pull-back of the tangent map $\partial_g \beta$, (note we are taking the pull-back of β viewed as a map

[24] Here we use Maurer-Cartan equation $0 = \mathrm{d}\theta + \frac{1}{2}[\theta \wedge \theta] = e_k(1) \otimes \left(\mathrm{d}\theta^k + \frac{1}{2} c^k_{ij} \theta^i \wedge \theta^j\right)$, which states θ is a flat gauge-field.

[25] Technically the HMC on compact semisimple matrix groups was constructed on the trivial bundle $\mathcal{G} \times \mathfrak{g}^*$, and although the momentum coordinates were left undefined, we can recover the mechanics therein by defining $\tilde{p}_i : \mathcal{G} \times \mathfrak{g}^* \to \mathbb{R}$ s.t. $\tilde{p}_i : (g, \alpha) \to \alpha(\xi_i)$, and constructing the symplectic structure $\tilde{\omega} = -\mathrm{d}(\tilde{p}_i \pi^* \theta^i)$. Clearly ω and $\tilde{\omega}$ are symplectomorphic with respect to the canonical isomorphism $T^*\mathcal{G} \cong \mathcal{G} \times \mathfrak{g}^*$, $(g, \alpha_g) \mapsto (g, L_g^* \alpha_g)$ (see §A.5.7.3).

$\mathcal{G} \to T^*\mathcal{G}$)

$$(\beta)^*\eta(g) \equiv \big(\partial_g\beta\big)^*\eta_{\beta(g)} = \eta_{\beta(g)} \circ \partial_g\beta = (p_i\pi^*\theta^i)_{\beta(g)} \circ \partial_g\beta.$$

We can write $\beta = f_i\theta^i$ for some functions $f_i \in C^\infty(\mathcal{G})$. It follows that $p\big(\beta(g)\big) = L_g^*\beta_g = L_g^*\big(f_i(g)\theta^i_g\big) = f_i(g)\theta^i(1)$, so $p_i\big(\beta(g)\big) = f_i(g)$, and, since $\pi \circ \beta : g \mapsto g$,

$$(\pi^*\theta^i)_{\beta(g)} \circ \partial_g\beta = \theta^i_g \circ \partial_{\beta(g)}\pi \circ \partial_g\beta = \theta^i_g \circ \partial_g(\pi \circ \beta) = \theta^i_g.$$

Hence $(\beta)^*\eta(g) = p_i\big(\beta(g)\big)(\pi^*\theta^i)_{\beta(g)} \circ \partial_g\beta = f_i(g)\theta^i_g = \beta(g)$. This proves that η is the canonical Liouville form (see proposition 3.2.11 of [2]). ■

In practice however we usually work in velocity space. Consider a Riemannian metric $\langle\cdot,\cdot\rangle$ on $T\mathcal{G}$. The associated musical isomorphism, enable us to pull-back ω to a symplectic structure $\omega_\flat$ on $T\mathcal{G}$ (recall §3.2.3). The latter can be expressed as the exterior derivative of a 1-form

$$\omega_\flat = -v^j\pi^*\big(\langle e_i, e_j\rangle\,\theta^i\big) \in \Gamma(T\mathcal{G}, T^*T\mathcal{G}),$$

where $v^j : T\mathcal{G} \to \mathbb{R}$ satisfies $(g, u_g) \mapsto \theta^i_g(u_g)$, and now $\pi : T\mathcal{G} \to \mathcal{G}$ is the tangent bundle projection (see §A.5.7.4). To take advantage of the symmetries of $\mathcal{G}$ we will restrict to left-invariant Riemannian metric, $\langle\cdot,\cdot\rangle \equiv G_{ij}\theta^i \otimes \theta^j$ where G_{ij} is symmetric positive definite, so the metric is determined by its value at the identity, $G_{ij} = \langle\xi_i, \xi_j\rangle_{\mathfrak{g}}$. Then

$$\omega_\flat \equiv -\mathrm{d}(G_{ij}v^j\pi^*\theta^i) = G_{ij}\pi^*\theta^i \wedge \mathrm{d}v^j + \tfrac{1}{2}G_{ir}v^r c^i_{jk}\pi^*\theta^j \wedge \pi^*\theta^k. \tag{3.20}$$

To simplify phase space we pullback $\omega_\flat$ to define an equivalent mechanics on $\mathcal{G} \times \mathfrak{g}$, which simply means the coordinate functions v^j are now the maps $(g, X) \mapsto \theta^j_1(X)$ (we will still denote the induced symplectic structure by $\omega_\flat$). In particular, if the metric is left-invariant the Hamiltonian $H \equiv V \circ \pi + \frac{1}{2}\langle\cdot,\cdot\rangle$ on $T\mathcal{G}$ defines an equivalent Hamiltonian $V \circ \pi + \frac{1}{2}\langle\cdot,\cdot\rangle_1$ on $\mathcal{G} \times \mathfrak{g}$ (see §A.5.7.5). Finally note the components of the curvature of a left-invariant metric are determined by the structure constants (see appendix B [97])

$$\Gamma_{ijk} = \tfrac{1}{2}\big(c_{ikj} - c_{jki} - c_{kji}\big), \qquad R^i_{jkl} = -\tfrac{1}{4}c^i_{mj}c^m_{kl},$$

and the symplectic volume form is given by the product of the Haar measure $\theta^1 \wedge \cdots \wedge \theta^n$ on $\mathcal{G}$ with the Lebesgue measure on the Lie algebra

$$\omega_\flat^n \equiv \bigwedge_{i=1}^n \omega_\flat \propto \underbrace{\mathrm{d}v^1 \wedge \cdots \wedge \mathrm{d}v^n}_{\text{Lebesgue on } \mathfrak{g}} \wedge \underbrace{\pi^*\theta^1 \wedge \cdots \wedge \pi^*\theta^n}_{\pi^*\text{Haar}}.$$

When $\mathcal{G} = \mathbb{R}^n$, $\theta^i = \mathrm{d}x^i$, and we recover the usual volume form.

3.9.1.2 Hamiltonian Fields and Euler-Arnold Equation

We now focus on the mechanical system $V + T$ on $\mathcal{G} \times \mathfrak{g}$ induced by a left-invariant metric $T = \frac{1}{2} v^i v^j G_{ij} : \mathfrak{g} \to \mathbb{R}$. For any function $f : \mathcal{G} \times \mathfrak{g} \to \mathbb{R}$, its Hamiltonian vector field $\hat{f} \in \Gamma(T\mathcal{G} \oplus T\mathfrak{g})$ is defined by $\omega_\flat(\hat{f}, \cdot) = \mathrm{d}f$. Following [187] but working in velocity space we have (see §A.5.7.6)

$$\hat{f} = G^{jk}\xi_k(f)e_j + \Big(G^{ik}G^{lj}v^r c_{rji}\xi_l(f) - G^{jk}e_j(f)\Big)\xi_k. \tag{3.21}$$

For the potential energy and kinetic energy induced by a left-invariant metric

$$\hat{T} = v^k e_k + G^{ik}v^r v^s c_{rsi}\xi_k \qquad \hat{V} = -G^{jk}e_j(V)\xi_k.$$

An integral curve γ_V of $\hat{V}$ satisfies $\dot{\gamma}_V = \hat{V}(\gamma)$. Let $\gamma_V(t) = \big(q(t), v(t)\big)$. The equation then reads $\big(\dot{q}, \dot{v}\big) = \big(0, -G^{jk}e_j(V)\xi_k\big)\big|_{\big(q(t),v(t)\big)}$, which implies

$$q(t) = q(0) \qquad v(t) = v(0) - tG^{jk}e_j(V)|_{q(0)}\xi_k. \tag{3.22}$$

For the integral curves of $\hat{T}$, we have $\big(\dot{q}, \dot{v}\big) = \big(v^k e_k, G^{ik}v^r v^s c_{rsi}\xi_k\big)|_{\gamma_T(t)}$. The mysterious term $G^{ik}v^r v^s c_{rsi}\xi_k$ disappeared in both [187, 186] (and did not even appear in [192]) since the structure constants of the Killing form are totally anti-symmetric (see §3.9.2.1). It is in fact nothing else than the $\langle\cdot,\cdot\rangle$-adjoint of the adjoint representation. Recall that the adjoint representation $\mathrm{ad} : \mathfrak{g} \to \mathrm{End}(\mathfrak{g})$ satisfies $\mathrm{ad}_\xi \equiv \mathrm{ad}(\xi) = [\xi, \cdot]$. For each $\xi \in \mathfrak{g}$ we can define its adjoint $\mathrm{ad}_\xi^\top$ with respect to the inner product $\big\langle \mathrm{ad}_\xi^\top X, Y\big\rangle = \langle X, \mathrm{ad}_\xi Y\rangle$. It follows that $\mathrm{ad}^\top$ is bilinear so

$$\begin{aligned}\big\langle \mathrm{ad}_v^\top v, \xi_k\big\rangle &= v^i v^j \big\langle \mathrm{ad}_{\xi_i}^\top \xi_j, \xi_k\big\rangle = v^i v^j \langle \xi_j, \mathrm{ad}_{\xi_i}\xi_k\rangle = v^i v^j \langle \xi_j, [\xi_i, \xi_k]\rangle \\ &= v^i v^j \langle \xi_j, c_{ik}^r \xi_r\rangle = c_{ik}^r v^i v^j g_{jr} = v^i v^j c_{jik}.\end{aligned}$$

On the other hand $\big\langle \mathrm{ad}_v^\top v, \xi_k\big\rangle = \big\langle \big(\mathrm{ad}_v^\top v\big)^s \xi_s, \xi_k\big\rangle = \big(\mathrm{ad}_v^\top v\big)^s G_{sk}$. Combining these two equations gives $\mathrm{ad}_v^\top v = G^{kl}v^i v^j c_{jik}\xi_l$. Hence the integral curves of $\hat{T}$ satisfy

$$\big(\dot{q}, \dot{v}\big) = \big(v^k e_k, \mathrm{ad}_v^\top v\big)|_{\gamma_T(t)}.$$

The equation $\dot{v} = \mathrm{ad}_v^\top v$ is a first order differential equation on the Lie algebra $\mathfrak{g}$ known as the Euler–Arnold equation. In order to derive the integral curves of $\hat{T}$ we can proceed in two steps. We first solve the Euler–Arnold equation to give a solution $s : [0, \ell] \to \mathfrak{g}$. The equation for $\dot{q}$

then states that $\dot{q} = v^k e_k|_{(q(t),s(t))} = s^k(t)e_k(q(t)) = \partial_1 L_{q(t)} s(t)$, which is a first order ODE for q called the reconstruction equation.

In summary we have shown that the Hamilton's equation $\dot{\gamma} = \hat{H}(\gamma)$ defined by the symplectic structure on $\mathcal{G} \times \mathfrak{g}$ reads

$$\dot{q} = v^j e_j(q) = \partial_1 L_q v \qquad \dot{v} = \mathrm{ad}_v^\top v - G^{jk} e_j(V)(q)\xi_k,$$

which can be recognised as the equations arising from the Euler–Poincaré reduction [163]. We finally note that the left-invariance of the metric enabled us to obtain two first order differential equations for the geodesic flow, rather than the usual second-order differential equation on a general Riemannian manifold.

3.9.1.3 Geodesic Integrators on Lie Groups

On an arbitrary Riemannian manifold $\mathcal{M}$, we can define the leapfrog integrator for the Hamiltonian system $H = V + T$ as a composition of the kinetic $\hat{T}$ and potential $\hat{V}$ fields defined by the canonical symplectic structure on $T\mathcal{M}$. The potential flow leads to the velocity update

$$q(t) = q(0) \qquad v(t) = v(0) - t\nabla_{q(0)} V, \tag{3.23}$$

and the kinetic (or geodesic) flow leads to the update $\big(q(t), v(t)\big) = \Phi_t^T\big(q(0), v(0)\big)$. Let us see the explicit relation between these updates, the ones derived in §3.9.1.2, and the ones used in lattice QCD. Suppose that $\mathcal{M} = \mathcal{G}$ is a Lie group and T is defined by a left-invariant metric $\langle \cdot, \cdot \rangle$. From §3.9.1.1 we know that, under the symplectomorphism $T\mathcal{G} \cong \mathcal{G} \times \mathfrak{g}$, the Hamiltonian vector-field induced by the inner product $\langle \cdot, \cdot \rangle_1 : \mathfrak{g} \to \mathbb{R}$ push-forwards to the geodesic flow induced by $T : T\mathcal{G} \to \mathbb{R}$. More precisely if $(q(t), v(t)) \in \mathcal{G} \times \mathfrak{g}$ is the integral curve of the Hamiltonian vector field of $\langle \cdot, \cdot \rangle_1$, then the geodesic flow of the Riemannian metric $T = \langle \cdot, \cdot \rangle$ on $T\mathcal{G}$ is given by $\big(q(t), \partial_1 L_{q(t)} v(t)\big) = \big(q(t), \dot{q}(t)\big)$. For the potential term, observe that[26]

$$\nabla_g V = G^{jk}(g)\mathrm{d}_g V(e_k) e_j(g) = G^{jk}(1) e_k(V)(g) \partial_1 L_g \xi_j$$

(by left-invariance G^{ij} is constant), and $v(t) = \theta_g^i(v(t)) e_i(g) = v^i(t)\partial_1 L_g \xi_i$. Thus under the symplectomorphism $\mathcal{G} \times \mathfrak{g} \to T\mathcal{G}$, we have $(\xi_i \to \partial_1 L_g \xi_i)$ and (3.23) push-forwards to the velocity update (3.22).

Finally let us see how the velocity update (3.22) is related to the ones used in lattice QCD, by considering the case when $\mathcal{G}$ is a matrix group (or we have an injective representation

[26] Note that if $\alpha \in \Gamma(T^*\mathcal{G})$, then $\alpha = e_i(\alpha)\theta^i = \alpha(e_i)\theta^i = \alpha_i \theta^i$. Moreover $u = G^{ij}\alpha_i e_j$ is the vector field associated to α by the metric since $\alpha(v) = \alpha_i \theta^i(v^j e_j) = \alpha_i v^i = \alpha_k \delta_i^k v^i = \alpha_k G^{kr} G_{ri} v^i = u^r G_{ri} v^i = \langle u, v \rangle$.

$\psi : \mathcal{G} \to \mathrm{GL}(n)$ s.t., $\partial_1\psi : \mathfrak{g} \to \mathbb{R}^{n\times n}$ is injective). In that case we need an extension W of V defined on an open subset of $\mathrm{GL}(n)$, in particular $W \circ \psi = V$. Then

$$e_i(V)(g) = \mathrm{d}(W \circ \psi)(e_i(g)) = \frac{\partial W}{\partial x^{jk}}\big(\partial\psi(e_i(g))\big)_{jk} = \mathrm{tr}\left(\frac{\partial W}{\partial x}^{\top} \cdot \psi(g) \cdot \partial_1\psi(\xi_i)\right),$$

where $\cdot$ is matrix multiplication, and we have used $\partial\psi(e_i(g)) = \psi(g) \cdot \partial_1\psi(\xi_i)$ [28]. If ψ is the inclusion $\iota : \mathcal{G} \to \mathrm{GL}(n)$, and we identify $g \sim \iota(g), \xi_i \sim \partial\iota(\xi_i)$, we recover §3.8.3. We arrive at the general velocity update formula for a matrix group (in lattice gauge theory we typically have $W(x) = \Re\,\mathrm{tr}\,(Ux)$ for a constant complex matrix U)

$$v^i(t)\xi_i = v^i(0)\xi_i - tG^{ij}(1)\,\mathrm{tr}\left(\frac{\partial W}{\partial x}^{\top} \cdot g \cdot \xi_j\right)\xi_i.$$

Given $H : \mathcal{G} \times \mathfrak{g} \to \mathbb{R}$, the following procedure 10 produces a sample from $e^{-H}\omega_\flat^n$; to obtain an HMC sample from the marginal $e^{-V}\theta^1 \wedge \cdots \wedge \theta^n$ we simply ignore the velocity component.

Algorithm 10 Algorithm to generate a sample from $e^{-H}\omega_\flat^n$, given current state $(q(0), v(0)) \in \mathcal{G} \times \mathfrak{g}$

1: sample $v \sim e^{-\frac{1}{2}G_{ij}v^jv^i}\mathrm{d}v = \mathcal{N}(0, G^{-1})$
2: $v \leftarrow v - \frac{1}{2}tG^{jk}\,\mathrm{tr}\left(\frac{\partial W}{\partial x}^{\top} \cdot q(0) \cdot \xi_j\right)\xi_k$
3: Solve the Euler–Arnold and Reconstruction equations $(q, v) \leftarrow (q(0), v)$
4: $v \leftarrow v - \frac{1}{2}tG^{jk}\,\mathrm{tr}\left(\frac{\partial W}{\partial x}^{\top} \cdot q \cdot \xi_j\right)\xi_k$
5: $\delta H \leftarrow H(q, v) - H(q(0), v(0))$
6: $U \sim \mathrm{Uniform}[0, 1]$
7: If $U < \exp(-\delta H)$, then return (q, v); else return $(q(0), -v(0))$

3.9.2 Geodesics

We now analyse how Euler–Arnold equation $\dot{v} = \mathrm{ad}_v^\top v$ simplifies when the inner product on $\mathfrak{g}$ is $\mathrm{ad}_\mathfrak{k}$-invariant for a subalgebra $\mathfrak{k} \subset \mathfrak{g}$ (defined below). When $\mathcal{G}$ is connected, $\mathrm{ad}_\mathfrak{k}$-invariant inner products correspond to left $\mathcal{G}$-invariant and right $\mathcal{K}$-invariant Riemannian metrics on $\mathcal{G}$.[27] When $\mathfrak{k} = \mathfrak{g}$ these correspond to bi-invariant Riemannian metric on $\mathcal{G}$, which always exists when $\mathcal{G}$ is compact. The results below extend some of the results from [13, 265, 35, 125, 253] concerning right $\mathcal{K}$-invariant Riemannian metrics on matrix groups and naturally reductive manifolds, with derivations directly from Euler–Arnold equation.

[27] i.e., the left and right actions L_g and R_k are isometries for $g \in \mathcal{G}, k \in \mathcal{K}$ (here $\mathcal{K}$ is the Lie group associated to $\mathfrak{k}$). When $\mathcal{G}$ is not connected an analogous correspondence exists with $\mathrm{Ad}_\mathcal{G}$ $\mathcal{K}$-invariant inner products, i.e., inner products s.t. $\langle \mathrm{Ad}_k\, u, \mathrm{Ad}_k\, v\rangle_1 = \langle u, v\rangle_1$ for $u, v \in \mathfrak{g}$, as we mentioned previously, and these are automatically $\mathrm{ad}_\mathfrak{k}$-invariant.

3.9.2.1 Ad $\mathcal{G}$-invariant Inner Products

The simplest case arises when $\operatorname{ad}_v^\top v = 0$. This situation occurs for example when the inner product on $\mathfrak{g}$ is ad-invariant, $\langle \operatorname{ad}_u v, w\rangle + \langle v, \operatorname{ad}_u w\rangle = 0$, which implies $\operatorname{ad}^\top = -\operatorname{ad}$ and $\operatorname{ad}_v^\top v = -[v,v] = 0$. We can find ad-invariant inner products on $\mathfrak{g}$ when $\mathcal{G}$ has an $\operatorname{Ad}\mathcal{G}$-invariant inner product, or equivalently if it has a bi-invariant metric. A particular $\operatorname{Ad}\mathcal{G}$-invariant inner product is available on Lie groups whose lie algebra is compact, which means the Killing-form $B(u,v) \equiv \operatorname{tr}\big(\operatorname{ad}(u)\operatorname{ad}(v)\big)$ is negative definite, and so $-B$ defines a positive definite ad-invariant inner product.

When $\operatorname{ad}_v^\top v = 0$, then $v(t) = v_0$, so the reconstruction equation becomes $\dot{q} = v^k e_k|_{(q(t),v_0)} = v_0^k e_k(q(t)) = \partial_1 L_{q(t)} v_0 = \frac{\mathrm{d}}{\mathrm{d}t} L_{q(t)} e^{tv_0}|_{t=0}$, which implies $q(t) = q(0)e^{tv_0}$ (here we used the fact that e^{tv_0} is tangent to v_0 at $t=0$). Hence we obtain the kinetic flow as in [187]

$$\gamma_T(t) = \big(q(0)e^{tv_0}, v_0\big).$$

3.9.2.2 Reductive Decomposition

Sometimes it is not possible to find a bi-invariant metric on $\mathcal{G}$. The natural next step is to consider inner products on $\mathfrak{g}$ which are $\operatorname{ad}_{\mathfrak{k}}$-invariant, where $\mathfrak{k}$ is a subalgebra of $\mathfrak{g}$. This means that $\langle \operatorname{ad}_u v, w\rangle + \langle v, \operatorname{ad}_u w\rangle = 0$ for all $u \in \mathfrak{k}$, and it follows that $\operatorname{ad}_u^\top = -\operatorname{ad}_u$. Then $\mathcal{K}$ is totally geodesic in $\mathcal{G}$ and its geodesics are given by 1-parameter subgroups (see §3.9.3.1). Let $\mathfrak{p} \equiv \mathfrak{k}^\perp$, so $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ and $v = v_{\mathfrak{k}} + v_{\mathfrak{p}}$. The Euler-Arnold equation reads $\dot{v} = \dot{v}_{\mathfrak{k}} + \dot{v}_{\mathfrak{p}} = \operatorname{ad}_v^\top v = \operatorname{ad}_{v_{\mathfrak{k}}}^\top v_{\mathfrak{k}} + \operatorname{ad}_{v_{\mathfrak{k}}}^\top v_{\mathfrak{p}} + \operatorname{ad}_{v_{\mathfrak{p}}}^\top v_{\mathfrak{k}} + \operatorname{ad}_{v_{\mathfrak{p}}}^\top v_{\mathfrak{p}} = -[v_{\mathfrak{k}}, v_{\mathfrak{p}}] + \operatorname{ad}_{v_{\mathfrak{p}}}^\top v_{\mathfrak{k}} + \operatorname{ad}_{v_{\mathfrak{p}}}^\top v_{\mathfrak{p}}$. We can simplify this further if the decomposition $\mathfrak{k} \oplus \mathfrak{p}$ is naturally reductive, i.e., $[\mathfrak{k}, \mathfrak{p}] \subset \mathfrak{p}$ and for all $u, v, w \in \mathfrak{p}$, $\langle [u,v]_{\mathfrak{p}}, w\rangle = \langle u, [v,w]_{\mathfrak{p}}\rangle$ [10]. This implies $\operatorname{ad}_{v_{\mathfrak{p}}}^\top v_{\mathfrak{p}} = 0$ (see §3.9.3.2), and we are left with

$$\dot{v} = -[v_{\mathfrak{k}}, v_{\mathfrak{p}}] + \operatorname{ad}_{v_{\mathfrak{p}}}^\top v_{\mathfrak{k}}.$$

Notably the constant curve $v = v_0$ with either $v_0 \in \mathfrak{k}$ or $v_0 \in \mathfrak{p}$ is a solution. We have thus shown directly from Euler-Arnold equations the result assumed in §3.1.3.1 that the geodesic starting with initial velocity in $\mathfrak{p}$ are given by matrix exponentiation (and the geodesic on the quotient space with the quotient metric are given by their projection).

3.9.2.3 Matrix Groups

Consider a subgroup $\mathcal{G} \subset \mathrm{GL}(n)$ and set $\mathfrak{g} = \mathfrak{p} \oplus \mathfrak{k}$ where $\mathfrak{p}$ is a subspace of the space of symmetric matrices, while $\mathfrak{k} = \mathfrak{so}(n)$. We have in mind the cases $\mathrm{GL}(n), SL(n)$ and $SO(n)$

with $\mathfrak{p}$ the space of symmetric matrices, the space of traceless symmetric matrices, and $\{0\}$ respectively. We equip $\mathfrak{g}$ with the standard inner product $\langle A, B\rangle \equiv \operatorname{tr}(A^\top B)$ which is $\operatorname{Ad}\mathcal{K}$-invariant. Since the product of a symmetric and an antisymmetric matrix is traceless, we have $\langle A, B\rangle = \operatorname{tr}(A_\mathfrak{p} B_\mathfrak{p}) - \operatorname{tr}(A_\mathfrak{k} B_\mathfrak{k})$. Note that for $\mathcal{G} = SO(n)$, this is just the negative of the Killing form (up to a positive constant); while on $\mathrm{GL}(n)$ and $SL(n)$, the Killing form, respectively given by $2n\operatorname{tr}(AB) - 2\operatorname{tr}(A)\operatorname{tr}(B)$ and $2n\operatorname{tr}(AB)$, doesn't define an inner product (it is degenerate and pseudo-Riemannian respectively [266]). For our inner product we have $\operatorname{ad}_S^\top A = [S, A]$ for any $S \in \mathfrak{p}, A \in \mathfrak{k}$, hence Euler-Arnold equation simplifies to $\dot{v} = -2[v_\mathfrak{k}, v_\mathfrak{p}]$ with solution (see §3.9.3.3)

$$\big(q(t), v(t)\big) = \Big(q(0)e^{\left(v_\mathfrak{p}(0) - v_\mathfrak{k}(0)\right)t} e^{2v_\mathfrak{k}(0)t}, v_\mathfrak{k}(0) + \operatorname{Ad} e^{-2v_\mathfrak{k}(0)t}(v_\mathfrak{p}(0))\Big). \tag{3.24}$$

3.9.2.4 Semi-simple Symmetric Spaces

The equations for the geodesics above apply in particular to semi-simple Lie groups $\mathcal{G}$ using an inner product based on the Killing form B (note $\mathrm{GL}(n)$ is not semi-simple). Let us consider a subgroup $\mathcal{K}$ for which $\mathcal{G}/\mathcal{K}$ is a symmetric space. In that case we can find a Cartan decomposition $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$. By corollary 5.4.3 [180] the Killing form is negative-definite on $\mathcal{K}$. We say $\mathcal{G}/\mathcal{K}$ is of compact (non-compact) type if B is negative (positive) definite on $\mathfrak{p}$. If $\mathcal{G}/\mathcal{K}$ is of compact type the inner product $\langle v, u\rangle \equiv B(v_\mathfrak{k}, u_\mathfrak{k}) + B(v_\mathfrak{p}, u_\mathfrak{p})$ is $\operatorname{Ad}\mathcal{G}$-invariant, while if it is of non-compact type $\langle v, u\rangle \equiv -B(v_\mathfrak{k}, u_\mathfrak{k}) + B(v_\mathfrak{p}, u_\mathfrak{p})$ is $\operatorname{Ad}\mathcal{K}$-invariant. Then $\operatorname{ad}_{v_\mathfrak{p}}^\top v_\mathfrak{k} = [v_\mathfrak{p}, v_\mathfrak{k}]$ (see Lemma 5.5.4 [180]) and we can find the integral curves of $\hat{T}$ as in §3.9.2.3.

3.9.3 Proofs

3.9.3.1 Totally Geodesic Subgroups

Recall we call a subgroup $\mathcal{K} \subset \mathcal{G}$ totally geodesic, if the geodesics of $\mathcal{K}$ are geodesics of $\mathcal{G}$. Suppose we have an $\operatorname{ad}_\mathfrak{k}$-invariant inner product on $\mathfrak{g}$. Let $\mathfrak{p} \equiv \mathfrak{k}^\perp$, then for any $u \in \mathfrak{k}, v \in \mathfrak{p}$, $\langle u, \operatorname{ad}_u v\rangle = -\langle \operatorname{ad}_u u, v\rangle = -\langle [u, u], v\rangle = 0$, so Theorem 2 [269] is satisfied. In fact $\mathfrak{k}$ is "easy totally geodesic" in $\mathfrak{g}$, i.e., $\operatorname{ad}_\mathfrak{k} \mathfrak{p} \subset \mathfrak{p}$, since for any $u, w \in \mathfrak{k}$, $v \in \mathfrak{p}$ we find $\langle \operatorname{ad}_u v, w\rangle = \langle v, \operatorname{ad}_u^\top w\rangle = -\langle v, [w, u]\rangle = 0$ since $[w, u] \in \mathfrak{k}$ and $v \in \mathfrak{p}$, hence $\operatorname{ad}_u v \in \mathfrak{p}$ (see def 4.1 [269]). Finally we note that the restriction of the inner product to $\mathfrak{k}$ is ad-invariant and thus the geodesics are given by the 1-parameter subgroups [269].

3.9.3.2 Naturally Reductive Geodesics

Recall that we have an $\mathrm{ad}_{\mathfrak{k}}$-invariant inner product, i.e. $\langle \mathrm{ad}_u\, v, w\rangle + \langle v, \mathrm{ad}_u\, w\rangle = 0$ for all $u \in \mathfrak{k}$, and a naturally reductive decomposition implies for $u, v, w \in \mathfrak{p}$, $\langle [u,v]_{\mathfrak{p}}, w\rangle = \langle u, [v,w]_{\mathfrak{p}}\rangle$, where $\mathfrak{p} = \mathfrak{k}^{\perp}$.

We want to show that $\mathrm{ad}_v^{\top}\, v = 0$ for any $v \in \mathfrak{p}$. Let $z \in \mathfrak{g}$. Then

$$\begin{aligned}\left\langle \mathrm{ad}_v^{\top}\, v, z\right\rangle &= \left\langle \mathrm{ad}_v^{\top}\, v, z_{\mathfrak{k}}\right\rangle + \left\langle \mathrm{ad}_v^{\top}\, v, z_{\mathfrak{p}}\right\rangle = \langle v, [v, z_{\mathfrak{k}}]\rangle + \langle v, [v, z_{\mathfrak{p}}]\rangle \\ &= \langle v, [v, z_{\mathfrak{k}}]\rangle + \langle [v,v], z_{\mathfrak{p}}\rangle = \langle v, [v, z_{\mathfrak{k}}]\rangle = -\left\langle v, \mathrm{ad}_{z_{\mathfrak{k}}}\, v\right\rangle,\end{aligned}$$

where we have used

$$\langle v, [v, z_{\mathfrak{p}}]\rangle = \langle v, [v, z_{\mathfrak{p}}]_{\mathfrak{p}}\rangle = \langle [v,v]_{\mathfrak{p}}, z_{\mathfrak{p}}\rangle = \langle [v,v], z_{\mathfrak{p}}\rangle = 0,$$

by orthogonality and the fact $[v,v] = 0$. Moreover $\langle v, \mathrm{ad}_{z_{\mathfrak{k}}}\, v\rangle = -\langle \mathrm{ad}_{z_{\mathfrak{k}}}\, v, v\rangle = -\langle v, \mathrm{ad}_{z_{\mathfrak{k}}}\, v\rangle$, so $\langle v, \mathrm{ad}_{z_{\mathfrak{k}}}\, v\rangle = 0$. In fact this only requires the inner product on $\mathfrak{g}$ to be $\mathrm{ad}_{\mathfrak{k}}\,\mathfrak{p}$-invariant, that is $\langle \mathrm{ad}_u\, v, w\rangle + \langle v, \mathrm{ad}_u\, w\rangle = 0$ for all $u \in \mathfrak{k}, v, w \in \mathfrak{p}$. Finally let us consider the term $\mathrm{ad}_{v_{\mathfrak{p}}}^{\top}\, v_{\mathfrak{k}}$. Note let S_i, A_j basis of $\mathfrak{p}, \mathfrak{k}$ respectively, and T_k arbitrary basis element of $\mathfrak{g}$. Then $\left\langle \mathrm{ad}_{S_i}^{\top}\, A_j, T_k\right\rangle = \langle A_j, [S_i, T_k]\rangle$. This is zero if $T_k \in \mathfrak{k}$, since $[\mathfrak{p}, \mathfrak{k}] \subset \mathfrak{p}$ and $\mathfrak{p} = \mathfrak{k}^{\perp}$. Moreover $\left\langle \mathrm{ad}_{S_i}^{\top}\, A_j, T_k\right\rangle = \langle A_j, c_{ik}^r T_r\rangle = c_{ik}^r g_{jr} = c_{jik}$ where $j \in \mathfrak{k}$, $i \in \mathfrak{p}$. On the other hand $\langle [S_i, A_j], T_k\rangle$ is again zero if $T_k \in \mathfrak{k}$; and $\langle [S_i, A_j], T_k\rangle = c_{ij}^r g_{rk} = c_{kij} = -c_{kji} = c_{ijk}$ for $i \in \mathfrak{p}$, $j \in \mathfrak{k}$, where in the last step we used $\mathrm{ad}_{\mathfrak{k}}$-invariance. It doesn't seem to be possible in general to relate c_{jik} with c_{ijk}.

3.9.3.3 Geodesics Matrix Groups

Let us first verify that $\mathrm{ad}_S^{\top}\, A = [S, A]$, for $S \in \mathfrak{p}$, $A \in \mathfrak{k}$. Let $T \in \mathfrak{g}$. Then

$$\begin{aligned}\left\langle \mathrm{ad}_S^{\top}\, A, T\right\rangle &= \langle A, [S,T]\rangle = \mathrm{tr}\left(A^T [S,T]\right) = -\,\mathrm{tr}\left(A[S,T]\right) \\ &= -\,\mathrm{tr}\left(AST\right) + \mathrm{tr}\left(ATS\right) = -\,\mathrm{tr}\left(AST\right) + \mathrm{tr}\left(SAT\right) \\ &= \mathrm{tr}\left([S,A]T\right) = \langle [S,A], T\rangle\end{aligned}$$

and since T is arbitrary the result holds. Here we have used the property of the trace that $\mathrm{tr}(ABC) = \mathrm{tr}(BCA)$ and the fact that $[S,A]$ is symmetric, $[S,A]^T = \left(SA - AS\right)^T = -AS + SA = [S,A]$. It follows that the Euler-Arnold equation becomes $\dot{v} = \dot{v}_{\mathfrak{k}} + \dot{v}_{\mathfrak{p}} = -2[v_{\mathfrak{k}}, v_{\mathfrak{p}}]$. Recall $\mathrm{ad}_u = \partial_1 \mathrm{Ad}(u) = \frac{d}{dt}\mathrm{Ad}(e^{tu})|_{t=0}$ for any $u \in \mathfrak{g}$.

For any $u, w \in \mathfrak{g}$ the curve $t \mapsto \mathrm{Ad}(e^{tu})(w)$ is tangent to $[u, \mathrm{Ad}(e^{tu})(w)]$ at t, since (for

matrix groups) $\mathrm{Ad}(e^{tu})w = e^{tu}we^{-tu}$ and differentiating gives $ue^{tu}we^{-tu} - e^{tu}we^{-tu}u = u \cdot \mathrm{Ad}(e^{tu})w - \mathrm{Ad}(e^{tu})w \cdot u = [u, \mathrm{Ad}(e^{tu})w]$; in particular at $t = 0$ this is $[u, w]$. Moreover note that $\mathrm{Ad}\,\mathcal{K}(\mathfrak{p}) \subset \mathfrak{p}$ (i.e., the decomposition of $\mathfrak{g}$ is reductive). Hence if $v(0) = v_{\mathfrak{k}}(0) + v_{\mathfrak{p}}(0)$ and using $\mathrm{Ad}(1)(u) = u$, we find that the curve

$$v(t) = v_{\mathfrak{k}}(0) + \mathrm{Ad}\, e^{-2v_{\mathfrak{k}}(0)t}(v_{\mathfrak{p}}(0))$$

satisfies the Euler-Arnold equation $\big($i.e. $v_{\mathfrak{k}}(t) = v_{\mathfrak{k}}(0)$ and $v_{\mathfrak{p}}(t) = \mathrm{Ad}\, e^{-2v_{\mathfrak{k}}(0)t}(v_{\mathfrak{p}}(0))\big)$. The reconstruction equation then implies [64]

$$q(t) = q(0)e^{(v_{\mathfrak{p}} - v_{\mathfrak{k}})t}e^{2v_{\mathfrak{k}}t}.$$

Note we can also write

$$v_{\mathfrak{p}} - v_{\mathfrak{k}} = \tfrac{1}{2}(v + v^T) - \tfrac{1}{2}(v - v^T) = v^T, \qquad 2v_{\mathfrak{k}} = v - v^T.$$

3.9.3.4 Force Integrators

Using (3.9.1.2), and the fact $\{r, h\} = \omega_{\flat}(\hat{r}, \hat{h}) = \hat{h}[r]$, we find

$$\begin{aligned}\{r, h\} &= G^{jk}\xi_k(h)e_j(r) + \Big(G^{ik}G^{lj}v^r c_{rji}\xi_l(h) - G^{jk}e_j(h)\Big)\xi_k(r)\\ &= G^{jk}\big(\xi_k(h)e_j(r) - e_j(h)\xi_k(r)\big) + v^r c_r^{lk}\xi_l(h)\xi_k(r)\end{aligned}$$

where $c_r^{lk} \equiv G^{ik}G^{lj}c_{rji}$. In particular

$$\{V, T\} = G^{jk}\xi_k(T)e_j(V) = G^{jk}G_{kr}v^r e_j(V) = v^r e_r(V)$$

and $\{V, \{V, T\}\} = G^{jk}\xi_k\big(v^r e_r(V)\big)e_j(V) = G^{jk}e_k(V)e_j(V)$. From (3.9.1.2) we find

$$\widehat{\{V, \{V, T\}\}} = -G^{jk}G^{ls}e_j\big(e_l(V)e_s(V)\big)\xi_k = -2G^{jk}G^{ls}e_l(V)e_j e_s(V)\xi_k$$

where we use Leibniz Rule. To see that $-2G^{jk}G^{ls}e_l(V)e_j e_s(V)\xi_k \in \mathfrak{p}$, we first observe that $e_j e_s(V) = [e_j, e_s](V) - e_s e_j(V)$. Moreover we can restrict to $s \in \mathfrak{p}$ since otherwise $e_s(V) = 0$ (note here the basis fields (e_i) are generated from an orthogonal decomposition $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$). Then $e_j(V) = 0$ for all $j \in \mathfrak{k}$, and using the properties of the stucture constant we have $[e_j, e_s] = c_{js}^r e_r$. We have $r \in \mathfrak{p}$ (again we have $e_r(V) = 0$), and note

$$G^{jk}G^{ls}e_l(V)[e_j, e_s](V)\xi_k = G^{jk}G^{ls}e_l(V)c_{js}^r e_r(V)\xi_k = G^{jk}G^{ls}G^{ra}e_l(V)c_{ajs}e_r(V)\xi_k$$

Hence, since $\mathfrak{p} = \mathfrak{k}^{\perp}$, $k \in \mathfrak{k}$ implies $j \in \mathfrak{k}$, implies $c_{ajs} = -c_{sja}$ and since $G^{ls}G^{ra}e_l(V)e_r(V)$ is symmetric in a, s the term sums to 0. We have thus shown that $\{V, \widehat{\{V, T\}}\} \in \mathfrak{p}$.

Chapter 4

The Bracket Geometry of Measure-Preserving Systems, and the Foundations of HMC

In §3 we have analysed in detail Hamiltonian Monte Carlo, which combines a deterministic proposal constructed from Hamiltonian dynamics, and a heat bath, typically a full momentum Gaussian refreshment. We have seen that some of the key properties of HMC are: **(i)** it is applicable to target densities known up to normalisation, **(ii)** the deterministic update is volume-preserving, and thus no Jacobian appears in the MC step, **(iii)** for symplectic integrators there exists a shadow Hamiltonian which ensures the acceptance rate remains high so that we can avoid random walks. We further observed that on Euclidean space, Hamiltonian mechanics could also be considered as a special case of the (underdamped) Langevin process (3.9), which motivated the irreversible extensions of HMC (see §3.4.2.3). In this chapter, following ideas from Volterra, Poincaré, de Rham, Koszul, Dufour, and Weinstein, we will develop the geometry of smooth measures, and characterise at the highest level of generality the classes of dynamical systems, mechanical systems, and diffusions that preserve a given smooth positive measure on arbitrary manifolds. Better, we will establish a fundamental connection between geometry and statistics, and as an immediate consequence obtain a canonical and complete recipe of measure-constraints preserving flows, as well as a closely related complete recipe of measure-preserving Nambu mechanics. We will also propose a generalised bracket MCMC algorithm, and understand exactly how the properties of Hamiltonian mechanics translate to properties of HMC, thus motivating several extensions.

We will begin by recollecting some elementary facts about diffusions on Euclidean space in

§4.1, and their applications to Markov chains. In §4.2 we will study the theory of smooth measures on manifolds, which are measures that assign zero weight to null sets of the manifold. Non-vanishing smooth measures, such as the target P, generate on manifolds considerable geometric structure, in particular a P-musical isomorphism $P^\flat$ between multi-vector fields and twisted differential forms, which in turn defines a derivative, the rotationnel curl_P. This shows that the standard operators from vector calculus div, curl naturally extend to manifolds without relying on a Riemannian metric: a measure provides the necessary structure. In particular, we can extend their vector calculus which allows us to characterise measure-preserving dynamical systems. We will also introduce brackets, linear differential operators on functions, that provide a general formalism to unify vector spaces of dynamical systems that depend on the target density only up to normalisation, such as gradient flows and Poisson mechanics. We will then consider applications of this bracket geometry. In §4.3 we will analyse key properties of HMC from the perspective of bracket geometries, to understand which generalised mechanical systems preserve some of the key benefits of HMC. Therein we will explain the important role played by the Jacobi identity, and introduce the differential operator $\phi_P^{\mathcal{B}}$, a generalisation of the modular vector field [358], which describes the compatibility between the mechanics of the bracket $\mathcal{B}$ and the measure P, and by construction vanishes iff P is an invariant "volume measure" of the mechanics of $\mathcal{B}$. Importantly, we will characterise measure-preserving $\mathcal{B}$-mechanics: these are locally the P-rotationnel of a 3-vector field. On Poisson manifolds the existence of an invariant measure is fully characterised by the modular class, and we will show that unimodular Poisson systems retain all the key properties of HMC in terms of implementation via splitting integrators, in particular the existence of a shadow Hamiltonian. Finally we will re-interpret the properties of HMC from the point of view of brackets, and after having generalised the "momentum flip" to arbitrary brackets, we will derive a bracket MCMC scheme generalising [114].

In §4.4 we will generalise the deterministic setting and study diffusions from the point of view of brackets. Most importantly, we will fully characterise measure-preserving diffusions. We will show that any measure-preserving diffusion is locally an $\mathcal{A}$-diffusion, generated by an antisymmetric bracket $\mathcal{A}$ (how "locally" depends on the de Rham cohomology). The result offers a constructive derivation that improves the complete recipe of [242] on $\mathbb{R}^n$, and most importantly extends it to arbitrary manifolds, relying on no other structure than the target measure. The proof in [242] relied on Fourier transforms and thus could not be extended to manifolds. But, as our proof shows, once we have understood the geometry of measures this result is in fact a direct consequence of Poincaré lemma, and the fact that the rotationnel and divergence coincide on vector fields. Indeed $\mathcal{A}$ may be thought as the "vector potential" of the Fokker-Planck "electromagnetic field". We shall analyse a symmetric-antisymmetric decomposition of these $\mathcal{A}$-diffusions on arbitrary manifolds, and show that their generators are

reversible up to appropriate "bracket-morphisms" - which generalises the corresponding result for Langevin diffusions. In §4.5 we will formulate the underdamped Langevin process on general manifolds, without relying on Lagrange multipliers as in [356, 329], and thus investigate Langevin-based irreversible HMC methods on non-Euclidean spaces. By explaining their relation to the Euler-Poincaré formalism we shall in particular generalise [17].

Finally, in §4.6 we will explain that the curl_P operator allows us to effectively construct a potential theory of measure-preserving flows, where the properties of the potential, for example its symmetries and the constraints it preserves, can be transferred to the induced field. This will lead us to the notion of Nambu mechanics, which provides us with an elegant interpretation in terms of flows of the P-musical isomorphisms: on n-twisted forms $P^\sharp$ is the Radon–Nikodym derivative, while for $\ell \geq 1$ it identifies $(n-\ell)$-twisted forms with ℓ-Nambu mechanics, a set of vector fields that preserve $(\ell-1)$ constraints. Moreover, closed $(n-\ell)$-twisted forms are identified with P-preserving ℓ-Nambu mechanics, and it turns out that, up to a topological obstruction belonging to the foliated de Rham cohomology, any flow that preserves a smooth positive measure and $(\ell-1)$-constraints is an exact ℓ-Nambu mechanics generated by those constraints. We will finish by briefly discussing the history of measure-preserving flows, noting that the Euclidean recipe of P-preserving flows [242] was already known in the SDE literature [140, 171], and explain why Hamiltonian mechanics is an excellent choice of measure-preserving flow in the context of sampling.

4.1 Diffusions on Euclidean Space

Let us start by considering time-homogeneous diffusions on $\mathbb{R}^n$ defined by stochastic differential equations (SDE) of the form

$$\mathrm{d}Z_t = b(Z_t)\mathrm{d}t + \sigma(Z_t)\mathrm{d}W_t, \tag{4.1}$$

where $b : \mathbb{R}^n \to \mathbb{R}^n$ is the **drift vector field**, W_t is a standard l-dimensional Brownian motion, and $\sigma : \mathbb{R}^n \to \mathbb{R}^{n\times l}$ is the **diffusion tensor**. Diffusions are **Markov processes** with no jumps. The drift and diffusion represent the rate of change of the mean and covariance[1] (2.5 [294])

$$b(z) = \lim_{t\to 0} \mathbb{E}\left[\frac{Z_{s+t} - Z_s}{t}\middle| Z_s = z\right],$$
$$\Sigma(z) \equiv \sigma(z)\sigma(z)^\top = \lim_{t\to 0} \mathbb{E}\left[\frac{(Z_{s+t} - Z_s)\otimes(Z_{s+t} - Z_s)}{t}\middle| Z_s = z\right].$$

[1] Recall the covariance of a random vector Z satisfies $\mathrm{cov}(Z,Z) = \mathbb{E}[Z\otimes Z] - \mathbb{E}[Z]\otimes\mathbb{E}[Z]$.

We are interested in choosing the coefficients b and σ in (4.1) to ensure the diffusion preserves the target distribution P, that is the triple (b, σ, P) satisfies Fokker–Planck equation (4.3). Note in this section "$\mathbb{R}^n$" may refer to either configuration space, in which the target is $P \propto e^{-V}\mathrm{d}q$, or phase space $\mathbb{R}^{\frac{1}{2}n} \times \mathbb{R}^{\frac{1}{2}n}$ (assuming n is even) where $P = \mu_H \propto e^{-H(q,p)}\mathrm{d}q\mathrm{d}p$.

From the transition kernel $\mathbb{P}_t(x \to \mathrm{d}y) \equiv \mathbb{P}[X_{t+s} \in \mathrm{d}y | X_s = x]$ we can define the **semigroup** (P_t) where P_t is a linear operator on the Banach space $(C_b(\mathbb{R}^n), \|\cdot\|_\infty)$

$$P_t f(x) \equiv \int f(y)\mathbb{P}_t(x \mapsto \mathrm{d}y),$$

which satisfies

$$P_{t+s} = P_t \circ P_s, \qquad P_0 f(x) = f(x).$$

The **generator** of (4.1) is defined as the limit (see 2.3 [294])

$$\mathcal{L}f \equiv \lim_{t \to 0} \frac{\|P_t f - f\|_\infty}{t},$$

where the domain of $\mathcal{L}$ is the subset of $C_b(\mathbb{R}^n)$ for which the above limit exists. If $\mathrm{d}z$ is the Lebesgue measure on $\mathbb{R}^n$, we denote by $\mathcal{L}^*$ the $L^2(\mathbb{R}^n, \mathrm{d}z)$-adjoint of $\mathcal{L}$

$$\langle \mathcal{L}^* f, h \rangle_{L^2} \equiv \langle f, \mathcal{L}h \rangle_{L^2} \qquad \text{i.e.,} \qquad \int (\mathcal{L}^* f) h \mathrm{d}z \equiv \int f(\mathcal{L}h)\mathrm{d}z.$$

For the diffusion (4.1) we have the explicit formulae for the generator and its adjoint (see (4.93)-(4.94) [294])

$$\mathcal{L} = b^i(z)\partial_{z^i} + \tfrac{1}{2}\Sigma^{jk}(z)\frac{\partial^2}{\partial z^j \partial z^k}, \quad \mathcal{L}^* f = \nabla \cdot \left(-b(z)f + \tfrac{1}{2}\nabla \cdot (\Sigma(z)f)\right). \tag{4.2}$$

If the distribution of the random vector Z_t satisfying (4.1) is given by $P_{Z_t} \equiv \rho_t \mathrm{d}z$ with initial condition $Z_0 = \rho_0 \mathrm{d}z$, the evolution of ρ_t is the solution $(z, t) \mapsto \lambda(z, t)$ of the **forward Kolmogorov equation**

$$\frac{\partial \lambda}{\partial t} = \mathcal{L}^* \lambda = \nabla \cdot \left(-b(z)\lambda + \tfrac{1}{2}\nabla \cdot (\Sigma(z)\lambda)\right), \qquad \lambda(\cdot, 0) = \rho_0.$$

The probability current J of a probability density ρ is defined as

$$J[\rho] \equiv -b(z)\rho + \tfrac{1}{2}\nabla \cdot (\Sigma(z)\rho).$$

In particular the density ρ of stationary distribution of the diffusion satisfies the divergence-free condition

$$0 = \nabla \cdot J[\rho] = \nabla \cdot \left(-b(z)\rho + \tfrac{1}{2}\nabla \cdot (\Sigma(z)\rho)\right) \tag{4.3}$$

which yields the equation that triple (b, σ, P) must satisfy. If the generator $\mathcal{L}$ is symmetric in $L^2(\mathbb{R}^n, P)$, then the density ρ (assumed positive) of the target P satisfies the stronger condition

$$J[\rho] = 0, \tag{4.4}$$

known as the **detailed balance condition**. The converse also holds, and these conditions are further equivalent to the process being P-reversible: $Z_0 \sim P$, then Z_t and Z_{T-t} have the same distribution (see proposition 4.5 [294]). For a general P-preserving diffusion (not necessarily satisfying detailed balance), the generator $\mathcal{L}$ decomposes into a symmetric and antisymmetric part (see (4.99) [294])

$$\mathcal{L}\cdot = \underbrace{\tfrac{1}{2}\frac{1}{\rho}\nabla\cdot(\rho\Sigma\nabla\cdot)}_{\text{symmetric part}} \quad + \quad \underbrace{\frac{1}{\rho}J[\rho]\cdot\nabla\cdot}_{\text{antisymmetric part}} \tag{4.5}$$

representing the reversible and irreversible components of the motion.

Given a target P, for many statistical problems (such as sampling) we are interested the coefficients (b, Σ) of the diffusion for which (4.3) holds. If we let Σ be an arbitrary semi-positive definite symmetric matrix, from the definition of the probability current we may write

$$b \equiv \tfrac{1}{2}\rho^{-1}\nabla\cdot(\rho\Sigma) + J$$

where J is a function s.t., $\nabla\cdot(J\rho) = 0$ (formally replacing $-J[\rho]/\rho$). If P is the unique stationary distribution of (4.1), and J is $\mathrm{d}z$-integrable, then theorem 2 [242] states there exists a skew-symmetric matrix field $z \mapsto A(z) \in \mathfrak{so}(n)$ s.t., the drift vector field takes the form

$$b = \tfrac{1}{2}\rho^{-1}\nabla\cdot(\rho(\Sigma + A)) = \tfrac{1}{2}\nabla\cdot(\Sigma + A) + \tfrac{1}{2}\rho^{-1}(\Sigma + A)\nabla\rho. \tag{4.6}$$

Conversely theorem 1 [242] states that if b is defined as in (4.6) where Σ is symmetric positive definite, and A is an arbitrary field of antisymmetric matrices, then (4.1) is ergodic and preserves P (the ergodicity claim is stated without proof, and seems wrong).

We next consider specifically Langevin processes which are central in molecular dynamics and statistics, and reduce to Hamiltonian mechanics in the noise-free case. We will start with the (underdamped) Langevin process in §4.1.1, and then move to overdamped Langevin systems in §4.1.2, mentioning their associated samplers in each case. We will later see in §4.5 that the common denominator between overdamped and underdamped Langevin diffusions is the absence of rotationnel-terms associated to the reference measure: Langevin processes are probability density based measure-preserving systems. In §A.4.10 we also introduce processes on discrete spaces, as these will be useful in §6.

4.1.1 Langevin Processes and Irreversible HMC

We focus now on the diffusions on phase space $\mathbb{R}^n \times \mathbb{R}^n$ which preserve the target measure $P = \mu_H \propto e^{-H(q,p)}\mathrm{d}q\mathrm{d}p$. In (4.6) we set

$$A \equiv -2J \equiv -2\begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}, \qquad \Sigma \equiv 2\begin{pmatrix} 0 & 0 \\ 0 & C \end{pmatrix}$$

for a symmetric matrix $C \in \mathbb{R}^{n\times n}$, so (4.1) becomes the **Langevin Process**

$$\mathrm{d}Q_t = \nabla_p H(Q_t, P_t)\mathrm{d}t, \tag{4.7a}$$
$$\mathrm{d}P_t = -\nabla_q H(Q_t, P_t)\mathrm{d}t - C\nabla_p H(Q_t, P_t)\mathrm{d}t + \sqrt{2C}\mathrm{d}W_t. \tag{4.7b}$$

As shown in 2.2.3.1 [329] the Langevin dynamics (4.7a) is ergodic provided the matrix C has full-rank, and is P-reversible up to momentum reversal $S(q,p) \equiv (q,-p)$ if the Hamiltonian is reversible, $H \circ S = H$, which means

$$\int f(\mathcal{L}h)\mathrm{d}P = \int (h \circ S)\mathcal{L}(f \circ S)\mathrm{d}P.$$

Note that if S were the identity the above would simply be the statement that $\mathcal{L}$ is symmetric in $L^2(\mathbb{R}^n \times \mathbb{R}^n, P)$ corresponding to the standard detailed balance condition.

For a standard "separable" Hamiltonian $H(q,p) \equiv \frac{1}{2}p^\top M^{-1}p + V(q)$ the Langevin process simplifies to

$$\mathrm{d}Q_t = M^{-1}P_t\mathrm{d}t, \qquad \mathrm{d}P_t = -\nabla_q V(Q_t)\mathrm{d}t - CM^{-1}P_t\mathrm{d}t + \sqrt{2C}\mathrm{d}W_t.$$

The generator then decomposes into a symmetric part associated to the Ornstein-Uhlenbeck (OU) process, and an antisymmetric part given by the deterministic Hamiltonian dynamics

$$\mathcal{L} = \underbrace{M^{-1}p \cdot \nabla_q \cdot - \nabla_q V \cdot \nabla_p \cdot}_{\text{antisymmetric Hamiltonian part}} + \underbrace{\nabla_p \cdot (C\nabla_q \cdot) - M^{-1}p \cdot C\nabla_p \cdot}_{\text{symmetric OU thermostat}}.$$

For the purpose of sampling it is natural to consider a discretisation of the Langevin process defined by composing a Hamiltonian part and a OU process as mentioned in §3.4.2.3, although more involved discretisations may be necessary when considering constrained processes (see 2.2.3.2 [329]).

4.1.2 MALA

Another important special case of (4.6) with $\Sigma = 2I$ and $A = 0$, so $b = \rho^{-1}\nabla\rho = \nabla \log \rho$, which gives the Smoluchowsky (P-reversible) diffusion, or overdamped Langevin process

$$\mathrm{d}Z_t = \nabla \log \rho(Z_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t,$$

which is ergocic by 2.2.2.1 [329]. The explicit Euler-Maruyama discretisation of the Smoluchowsky diffusion reads

$$Z_{n+1} = Z_n + \delta t \nabla \log \rho(Z_n) + \sqrt{2\delta t}\xi, \qquad \xi \sim \mathcal{N}(0, I)$$

The Z_{n+1} is thus distributed as

$$Z_{n+1} \sim \mathcal{N}\left(Z_n + \delta t \nabla \log \rho(Z_n), 2\delta t I\right).$$

The discretisation of the Smoluchowsky diffusion introduces an error in the dynamics which prevents the Markov chain from converging to P. In order to use the chain to produce samples from P, we can consider the above as a proposal kernel, $T(x \mapsto \mathrm{d}y) = \mathcal{N}\left(x + \delta t \nabla \log \rho(x), 2\delta t I\right)(\mathrm{d}y)$, and use a Metropolis accept-reject step as in 1. This is called the **Metropolis Adjusted Langevin Algorithm** (MALA). If we ignore the accept-reject step, and thus view T as the transition kernel $\mathbb{P}$, we call this the **Unadjusted Langeving Algorithm** (ULA). It is important to note that the discretisation also compromises the nice convergence properties that the continuous diffusion may have [307].

4.2 The Geometry and Dynamics Generated by Smooth Measures and the Bracket Formalism

Manifolds are topological spaces, and thus have a Borel algebra generated by the open sets of the topology, and the associated Borel measures are naturally compatible with the topology of the manifold. Better, while as mentioned in §3.1 there is no such thing as a canonical measure on a general manifold, there is on the other hand a well-defined notion of measure-zero sets generated by the charts [209, lemma 6.6]. In particular, any absolutely continuous Radon (signed) measure P, that is it assigns $|P|$-null weight to null sets, may be equivalently viewed as a (measurable) one-density, or equivalently as a top-form taking values in the orientation

bundle, see 10 [264]. In particular, as we show in §4.2.1, positive smooth measures determine a derivative, the P-rotationnel which adapts to the measure the exterior derivative, and allows us to transform Poincaré lemma into a theorem about the characterisation of P-preserving dynamical systems. In §4.2.2 we will then introduce brackets, and study general properties of their associated mechanical systems. As we will show later these brackets are the appropriate mathematical objects to describe many statistical algorithms, from HMC to Stein operators.

4.2.1 The Geometry and Dynamics of Measures

Consider a **volume manifold** $(\mathcal{M}, P)$ where $\mathcal{M}$ is an n-dimensional manifold, and P a smooth positive measure, that is a positive tensor density of weight one. Following ideas from Koszul and Dufour [197, 101] we introduce the musical isomorphisms and a derivative on multi-vector fields generated by the measure P:

Definition 2 (Rotationnel of a k-vector field)**.** *Given a volume manifold $(\mathcal{M}, P)$ of dimension n, the P-**rotationnel** of a k-vector field $curl_P : \Gamma(\Lambda^k T\mathcal{M}) \to \Gamma(\Lambda^{k-1} T\mathcal{M})$ for $k \leq n$ is defined as* [2]

$$curl_P(X) = P^\sharp \circ \mathrm{d} \circ P^\flat(X), \quad X \in \Gamma(\Lambda^k T\mathcal{M}),$$

where the operator $P^\flat : \Gamma(\Lambda^k T\mathcal{M}) \to \Gamma(\Lambda^{n-k} T^*\mathcal{M})$ is a $C^\infty(\mathcal{M})$-linear isomorphism defined as $P^\flat(X) \equiv i_X P$ and $P^\sharp$ is its inverse.[3]

The isomorphisms defined by P generalise an important map in statistics: when applied to a function f the map $P^\flat$ returns $P^\flat(f) = fP$, while $P^\sharp$ applied to a smooth measure Q yields its Radon-Nikodyn derivative with respect to P, $P^\sharp(Q) = \frac{\mathrm{d}Q}{\mathrm{d}P}$. In particular when applied to vector fields the P-rotationnel acts as the divergence operator $\mathrm{curl}_P = \mathrm{div}_P : \mathfrak{X}(\mathcal{M}) \to C^\infty(\mathcal{M})$, which follows from

$$\mathcal{L}_X P = \mathrm{d} i_X P = P^\flat \mathrm{curl}_P(X) = \mathrm{curl}_P(X) P,$$

together with the definition of $\mathrm{div}_P(X)$ as the function satisfying $\mathrm{div}_P(X)P = \mathcal{L}_X P$. Note also that when $\mathcal{M} = \mathbb{R}^3$ and $P = \mathrm{d}\mathbf{x} \equiv \mathrm{d}x\mathrm{d}y\mathrm{d}z$, setting $\mathcal{A} \equiv \mathcal{A}_x \partial_y \wedge \partial_z + \mathcal{A}_y \partial_z \wedge \partial_x + \mathcal{A}_z \partial_x \wedge \partial_y$ allows us to recover the usual vector calculus definition for the curl of the vector $(\mathcal{A}_x, \mathcal{A}_y, \mathcal{A}_y)$

$$\mathrm{curl}_{\mathrm{d}\mathbf{x}}(\mathcal{A}) = \left(\frac{\partial \mathcal{A}_y}{\partial z} - \frac{\partial \mathcal{A}_z}{\partial y} \right) \partial_x + \left(\frac{\partial \mathcal{A}_z}{\partial x} - \frac{\partial \mathcal{A}_x}{\partial z} \right) \partial_y + \left(\frac{\partial \mathcal{A}_x}{\partial y} - \frac{\partial \mathcal{A}_y}{\partial x} \right) \partial_z.$$

[2] When the manifold is orientable, P is a volume form, and everything makes sense. In the non-orientable case, we use the fact the space of one densities is isomorphic to the space of twisted n-forms (forms that take values in the orientation line bundle). Viewing P as a twisted n-form allows to naturally extend the operations of Lie derivative, exterior derivative, interior product... In particular Cartan's magic formula holds (8.2 A [3]), and $P^\flat$ becomes an isomorphism between multi-vector fields and twisted forms.

[3] Here $i_X P(A) \equiv \langle P, X \wedge A \rangle$ for any $A \in \Gamma(\Lambda^{n-k} T\mathcal{M})$, see 2.5 [102] (the pairing is well-defined since $(\Gamma(\Lambda^\ell T\mathcal{M}))^* \cong \Gamma(\Lambda^\ell T^*\mathcal{M})$).

More generally when $\mathcal{M}$ is Riemannian we typically identify $\Lambda^k T\mathcal{M} \cong \Lambda^k T^*\mathcal{M}$ using the musical isomorphisms in which case the map $\text{vol}^\flat$ associated to the Riemannian measure vol becomes the Hodge star $\star$, and the curl_{vol}, viewed as a map on k-forms, is called the co-differential, usually denoted $\delta \equiv \star^{-1}\text{d}\star = \sharp \circ \text{curl}_{\text{vol}} \circ \flat$. This shows that the standard operators div and curl from vector calculus are in fact generated by measures - precisely the object we are given in statistics! We mention that many authors define the curl of a "vector field" $X : \mathbb{R}^n \to \mathbb{R}^n$ as the field $\text{curl}(X) : \mathbb{R}^n \to \mathbb{R}^{\frac{1}{2}n(n-1)}$ with components $\partial_i X^j - \partial_j X^i$ (for $1 \leq i < j \leq n$), which simply corresponds to viewing X as a 1-form, and setting $\text{curl}(X) \equiv \text{d}X$. In that context, the statement that curl-free vector fields are the gradient of a function, is just the statement that closed 1-forms are exact.

When constructing dynamical-based sampling methods, we are interested in determining local flows Φ_t^X that are target-preserving (for each t in their domain of definition). Then given any current sample $z \sim P$, we specify a t and return a new sample $\Phi_t^X(z) \sim P$. Such local flows are generated by vector fields that are incompressible with respect to P, that is they are divergence-free. We will denote by $\text{Dyn}(P) \equiv \{X \in \mathfrak{X}(\mathcal{M}) : \mathcal{L}_X P = 0\}$ the set of P-preserving vector fields, which forms a Lie subalgebra of $(\mathfrak{X}(\mathcal{M}), [\cdot, \cdot])$ since $\mathcal{L}_{[X,Y]} = \mathcal{L}_X \circ \mathcal{L}_Y - \mathcal{L}_Y \circ \mathcal{L}_X$. It is worth mentioning that the completeness of the flow of P-preserving dynamics can be characterised (recall a vector field is complete if its integral curves exists for all times): $X \in \text{Dyn}(P)$ is a complete vector field P-a.e. iff iX, viewed as an operator on $L^2(\mathcal{M}, P)$, is essentially self-adjoint on $C_c^\infty(\mathcal{M})$ (8 [3]).

The P-rotationnel satisfies the important identity $\text{curl}_P \circ \text{curl}_P = P^\sharp \circ \text{d}^2 \circ P^\flat = 0$, i.e., it is a boundary operator and thus induces a chain complex and homology groups. It follows that the P-rotationnel of any $\mathcal{A} \in \Gamma(\Lambda^2 T\mathcal{M})$ defines a P-preserving dynamics,

$$\text{div}_P \text{curl}_P(\mathcal{A}) = 0,$$

which is the generalisation of the famous identity

$$\nabla \cdot \nabla \times = 0$$

to arbitrary volume manifolds. This key result provides a simple, yet general, recipe to construct target-preserving dynamics. We may then wonder whether the converse is true, that is,

$$\text{if } X \in \text{Dyn}(P) \text{ then } X = \text{curl}_P(\mathcal{A}) \text{ for some } \mathcal{A}.$$

This would be the generalisation of the standard identity from vector calculus on $\mathbb{R}^3$

$$\text{if } \nabla \cdot Y = 0, \text{ then there exists } A \text{ s.t., } Y = \nabla \times A.$$

More generally, this question relates to characterising the difference between the kernel of $\text{curl}_P|_{\mathfrak{X}^\ell(\mathcal{M})}$ and the image of curl_P on $\mathfrak{X}^{\ell+1}(\mathcal{M})$,[4] and this is precisely what the homology groups of the boundary operator curl_P describe:

$$\mathcal{H}_P^\ell(\mathcal{M}) \equiv \frac{\ker\left(\text{curl}_P : \mathfrak{X}^\ell(\mathcal{M}) \to \mathfrak{X}^{\ell-1}(\mathcal{M})\right)}{\text{Im}\left(\text{curl}_P : \mathfrak{X}^{\ell+1}(\mathcal{M}) \to \mathfrak{X}^\ell(\mathcal{M})\right)}.$$

The following fundamental result shows that the homology groups of P are isomorphic to the (twisted) de Rham cohomology groups, which allows us to characterise on arbitrary manifolds the space of curl_P-vanishing multi-vector fields. Moreover it shows that the curl_P operator naturally "commutes" with restrictions to open sets, and thus that P-preserving dynamics are equivalent to locally curled vector fields.

Theorem 3 (**The Homology of P**). *The isomorphism $P^\flat$ descends to an isomorphism between the homology groups $\mathcal{H}_P^\ell(\mathcal{M})$ of $curl_P$ and the twisted de Rham cohomology groups $H^{n-\ell}(\mathcal{M})$. Hence*

$$\ker\left(curl_P|_{\mathfrak{X}^\ell(\mathcal{M})}\right) \cong curl_P\left(\mathfrak{X}^{\ell+1}(\mathcal{M})\right) \oplus P^\sharp\left(H^{n-\ell}(\mathcal{M})\right).$$

In particular

$$Dyn(P) \cong curl_P\left(\mathfrak{X}^2(\mathcal{M})\right) \oplus P^\sharp\left(H^{n-1}(\mathcal{M})\right),$$

and any P-preserving dynamics will be globally the P-rotationnel of some $\mathcal{A} \in \mathfrak{X}^2(\mathcal{M})$ iff the $(n-1)$ de twisted Rham cohomology is trivial. Moreover, if $U \subset \mathcal{M}$ is an open subset, then

$$|_U \circ curl_P = curl_{P|_U} \circ |_U.$$

Hence the set of P-preserving dynamics is precisely the set of locally curled vector fields

$$Dyn(P) = Curl_{loc}(P)$$

where $Curl_{loc}(P) = \{X \in \mathfrak{X}(\mathcal{M}) : \forall q$, there is a neighbourhood U and $\mathcal{A} \in \mathfrak{X}^2(U)$ s.t., $X = curl_{P|_U}(\mathcal{A})\}$.

Proof. The P-derivative $\text{curl}_P \equiv P^\sharp \circ \text{d} \circ P^\flat : \mathfrak{X}^k(\mathcal{M}) \to \mathfrak{X}^{k-1}(\mathcal{M})$ is a vector space homomorphism, satisfying $\text{curl}_P \circ \text{curl}_P = 0$, that it is a boundary operator on the chain complex of k-multi-vector fields, and thus $\text{Im}\left(\text{curl}_P : \mathfrak{X}^{\ell+1}(\mathcal{M}) \to \mathfrak{X}^\ell(\mathcal{M})\right)$ is a linear subspace of $\ker\left(\text{curl}_P : \mathfrak{X}^\ell(\mathcal{M}) \to \mathfrak{X}^{\ell-1}(\mathcal{M})\right)$. We can then define the ℓ^{th} Holomology group

$$\mathcal{H}_P^\ell(\mathcal{M}) \equiv \frac{\ker\left(\text{curl}_P : \mathfrak{X}^\ell(\mathcal{M}) \to \mathfrak{X}^{\ell-1}(\mathcal{M})\right)}{\text{Im}\left(\text{curl}_P : \mathfrak{X}^{\ell+1}(\mathcal{M}) \to \mathfrak{X}^\ell(\mathcal{M})\right)},$$

[4] Here we have denoted the space of ℓ-vector fields by $\mathfrak{X}^\ell(\mathcal{M}) \equiv \Gamma(\Lambda^\ell T\mathcal{M})$.

and in particular the first one provides information on P-preserving vector fields (using $\mathrm{curl}_P(X) = \mathrm{div}_P(X)$)

$$\mathcal{H}^1_P(\mathcal{M}) \equiv \frac{\ker\left(\mathrm{div}_P : \mathfrak{X}(\mathcal{M}) \to C^\infty(\mathcal{M})\right)}{\mathrm{Im}\left(\mathrm{curl}_P : \mathfrak{X}^2(\mathcal{M}) \to \mathfrak{X}(\mathcal{M})\right)}.$$

Notice that the map $[\cdot] \circ P^\flat : \Gamma(\Lambda^\ell T\mathcal{M}) \to H^{n-\ell}(\mathcal{M})$ descends to a map $\mathcal{H}^\ell_P(\mathcal{M}) \to H^{n-\ell}(\mathcal{M})$, since $[\cdot] \circ P^\flat(X + \mathrm{curl}_P(Y)) = [P^\flat(X) + \mathrm{d}i_Y P] = [P^\flat(X)]$. The map is surjective since $P^\flat$ is, and injective since $[P^\flat(X)] = [P^\flat(Y)] \implies [P^\flat(X - Y)] = 0 \implies P^\flat(X - Y) = \mathrm{d}\alpha \implies X - Y = \mathrm{curl}_P(Z) \implies [X] = [Y]$. In particular, $H^{n-1}(\mathcal{M})$ is trivial iff $\mathcal{H}^1_P(\mathcal{M})$ is, in which case every divergence-free vector field is the curl_P of a bi-vector field $\mathcal{A}$.

In general, we can still use Poincaré lemma (or Volterra theorem, as it was proved by Vito Volterra [355]) and the properties of curl_P to show that any P-preserving vector field is locally a curl vector field. Denoting the inclusion by $\iota_U : U \hookrightarrow \mathcal{M}$, we have $\iota_U^*(P^\flat(Y)) = \iota_U^*(i_Y P) = i_{Y_U}\iota_U^* P = i_{Y_U} P_U = P_U^\flat(Y_U)$, where P_U, Y_U denote their restriction to U, and we have used proposition 7.4.10 [3]; hence $\iota_U^* \circ P^\flat = P_U^\flat \circ |_U$. Setting $Y = P^\sharp(\alpha)$ this yields $P^\sharp(\alpha)|_U = P_U^\sharp(\iota_U^*\alpha)$ for any twisted form α. Hence

$$|_U \circ \mathrm{curl}_P \equiv |_U \circ P^\sharp \circ \mathrm{d} \circ P^\flat = P_U^\sharp \circ \iota_U^* \circ \mathrm{d} \circ P^\flat = P_U^\sharp \circ \mathrm{d} \circ \iota_U^* \circ P^\flat = P_U^\sharp \circ \mathrm{d} \circ P_U^\flat = \mathrm{curl}_{P_U} \circ |_U.$$

Thus, $\mathrm{curl}_P(Y) = 0$ iff $P^\sharp \circ \mathrm{d} \circ P^\flat(Y) = 0$ iff $\mathrm{d} \circ P^\flat(Y) = 0$ (since $P^\sharp$ is a linear isomorphism). By Poincaré Lemma, this holds iff around any point there is an open neighbourhood U over which $P^\flat(Y)|_U = \mathrm{d}\alpha$ for some $(\dim(\mathcal{M}) - 2)$-form α on U. Then $\mathrm{d}\alpha = \iota_U^*(P^\flat(Y)) = P_U^\flat(Y_U)$, and $P_U^\flat(Y|_U) = \mathrm{d}\alpha$ iff $Y_U = P_U^\sharp \mathrm{d}\alpha = \mathrm{curl}_{P_U}(\mathcal{A})$ where $\mathcal{A} \equiv P_U^\sharp(\alpha)$ is a 2-vector on U.

Finally we also mention that when $P = \mathrm{d}x$ is the Lebesgue measure on Euclidean space, $\mathrm{Dyn}(\mathrm{d}x) = \mathrm{curl}_{\mathrm{d}x}\left(\mathfrak{X}^2(\mathcal{M})\right)$ was essentially already proved by Vito Volterra [355], that the statement $\mathrm{Dyn}(P) \cong \mathrm{curl}_P\left(\mathfrak{X}^2(\mathcal{M})\right) \oplus P^\sharp\left(H^{n-1}\left(\mathcal{M}\right)\right)$ appears in implicit form (essentially written as $P^\flat\left(\mathrm{Dyn}(P)\right) = \mathrm{d}\Omega^{n-2}(\mathcal{M}) \oplus H^{n-1}(\mathcal{M})$) in [260, Thm. 6], and that by Poincaré duality if $\mathcal{M}$ has a finite good cover (in which case the cohomology groups are finite dimensional) we may alternatively work with the first compactly supported de Rham cohomology group (proposition 5.3.1 and theorem 7.8 [52]). ■

Finally, observe that the measure P also defines a pairing $\langle \cdot, \cdot \rangle_P : \Gamma(\Lambda^k T^*\mathcal{M}) \times \Gamma(\Lambda^k T\mathcal{M}) \to \mathbb{R}$ between forms and multivector fields by integration which will be useful when we discuss the extension of Stein discrepancies to manifolds (see §6.10)

$$\langle \alpha, V \rangle_P \equiv \int \alpha \wedge i_V P. \tag{4.8}$$

Provided the boundary term vanishes, the formal adjoint of the exterior derivative with respect to the pairing is precisely the rotationnel, $\mathrm{d}^* = (-1)^\ell \mathrm{curl}_P$ on $\Gamma(\Lambda^\ell T\mathcal{M})$, since for

$\alpha \in \Gamma(\Lambda^k T^*\mathcal{M})$

$$\begin{aligned}\langle \mathrm{d}\alpha, V\rangle_P &= \int \mathrm{d}\alpha \wedge i_V P = \int \mathrm{d}(\alpha \wedge i_V P) - (-1)^k \int \alpha \wedge \mathrm{d} i_V P \\ &= (-1)^{k+1} \int \alpha \wedge \mathrm{d} i_V P = (-1)^{k+1} \int \alpha \wedge P^\flat \circ P^\sharp \mathrm{d} i_V P \\ &= (-1)^{k+1} \int \alpha \wedge P^\flat \mathrm{curl}_P(V) = (-1)^{k+1} \int \alpha \wedge i_{\mathrm{curl}_P(V)} P \\ &= (-1)^{k+1} \langle \alpha, \mathrm{curl}_P(V)\rangle_P .\end{aligned}$$

When $k = 0$ the pairing recovers the (pseudo-) inner product on square P-integrable functions $\langle f, h\rangle_P = \int fhP$, while when $k = 1$ we can write $\langle \alpha, X\rangle_P = \int \alpha(X)P$ [5]. When $P = \mathrm{vol}$ is the Riemannian measure, and we identify multivector fields with differential forms via the musical isomorphism as discussed above, we further recover the standard inner product between differential forms appearing the theory of Harmonic forms

$$(\alpha, \beta)_{\mathrm{vol}} = \int \alpha \wedge \star\beta. \tag{4.9}$$

4.2.2 The Bracket Formalism

Instances of brackets are central in many areas of physics, including classical mechanics, quantum physics, and fluid dynamics [205, 23, 271, 165]. In this section we develop the general theory of brackets, in particular we discuss the vector spaces of vector fields they generate and their canonical transformations. In the remaining sections of the rest of the chapter we will see how these brackets also provide key insights into statistical problems, and derive additional properties of brackets suited to the application of interest. In §4.2.2.1 we will introduce the concept of bracket and their associated vector space of dynamics. In §4.2.2.2 and §4.2.2.3 we will briefly discuss their canonical transformations and Casimirs. In the following we work on an arbitrary bundle $\pi : \mathcal{F} \to \mathcal{M}$ over the sample space $\mathcal{M}$, as in practice the $\mathcal{B}$-mechanics we discuss below are often used on a higher dimensional space (for example $\mathcal{F} = T^*\mathcal{M}$ in HMC).

4.2.2.1 $\mathcal{B}$-Hamiltonian Dynamics

Generalising the relation between Poisson brackets and Poisson bivector fields we have the following

Lemma 4. *There is a bijection between:*

[5] since $0 = i_X(\alpha \wedge P) = i_X \alpha P + (-1)^k \alpha \wedge i_X P$

1. *$C^\infty(\mathcal{F})$-linear maps $\mathcal{B}^\sharp : \Omega^1(\mathcal{F}) \to \mathfrak{X}(\mathcal{F})$,*

2. *$C^\infty(\mathcal{F})$-bilinear maps $\mathcal{B} : \Omega^1(\mathcal{F}) \times \Omega^1(\mathcal{F}) \to C^\infty(\mathcal{F})$, and*

3. *$\mathbb{R}$-bilinear maps $\{\cdot,\cdot\} : C^\infty(\mathcal{F}) \times C^\infty(\mathcal{F}) \to C^\infty(\mathcal{F})$ satisfying the Leibniz rule in each entry.*

Proof. (1) defines a $(0,2)$-tensor $\mathcal{B} : \Omega^1(\mathcal{F}) \times \Omega^1(\mathcal{F}) \to C^\infty(\mathcal{F})$ by $\mathcal{B}(\alpha,\beta) \equiv \langle \beta, \mathcal{B}^\sharp(\alpha) \rangle_*$ and vice-versa, thus proving (1) $\Leftrightarrow$ (2). Moreover a tensor $\mathcal{B}$ as in (2) can be identified with a "bracket": $\mathbb{R}$-bilinear map $\{\cdot,\cdot\} : C^\infty(\mathcal{F}) \times C^\infty(\mathcal{F}) \to C^\infty(\mathcal{F})$ satisfying the Leibniz rule in each entry by $\{f,g\} \equiv \mathcal{B}(\mathrm{d}f, \mathrm{d}g)$ (since $\mathrm{d}(fg) = g\mathrm{d}f + f\mathrm{d}g$). Conversely given a bracket, the $\mathbb{R}$-linearity and Leibniz rule implies that $\{f,g\}(z)$ only depends on $\mathrm{d}_z f$ and $\mathrm{d}_z g$ and that $\mathcal{B}_z(\mathrm{d}_z f, \mathrm{d}_z g) \equiv \{f,g\}(z)$ is $\mathbb{R}$-bilinear, which defines a tensor field $\mathcal{B} : z \mapsto \mathcal{B}_z$ satisfying (2), thus proving (2) $\Leftrightarrow$ (3). ■

The bracket in (3) is symmetric/alternating iff $\mathcal{B}$ in (2) is symmetric/alternating, and anti-symmetric brackets will be denoted by $\mathcal{A} \equiv \mathcal{B}$ throughout the thesis. Any bracket generates a mechanical system, that is a vector space of dynamics, by associating differentiable functions to generalised "Hamiltonian vector fields" as follows:

Definition 5 ($\mathcal{B}$-Hamiltonian vector field)**.** *Given a C^1 function f on a manifold $\mathcal{F}$, equipped with a* ***bracket*** *$\{\cdot,\cdot\}$ or its equivalent $(0,2)$-tensor field $\mathcal{B}$, the associated* ***$\mathcal{B}$-Hamiltonian vector field*** *X_f is defined by $X_f^\mathcal{B} \equiv \mathcal{B}^\sharp(\mathrm{d}f)$, or, equivalently, $X_f^\mathcal{B} g \equiv \{f,g\}$. The assignment $f \mapsto X_f^\mathcal{B}$ is $\mathbb{R}$-linear; we shall denote the vector space of $\mathcal{B}$-Hamiltonian vector fields over $\mathcal{F}$ by $\mathfrak{X}_\mathcal{B}(\mathcal{F})$, and refer to it as the* ***$\mathcal{B}$-mechanics****.*

We often drop the superscript, $X_f \equiv X_f^\mathcal{B}$, and as in HMC we will denote by $\Phi^f \equiv \Phi^{X_f}$ the flow of X_f. Locally, the $\mathcal{B}$-Hamiltonian vector field reads

$$X_f(z) = X_f^i(z)\partial_i = \mathcal{B}^{ji}(z)\partial_j f \partial_i$$

so the components of X_f are pointwise linearly related to the differential of f, and in particular only depend on the equivalence class $[f]$ of f, where $f \sim h$ iff $f = h + c$ with c a locally constant function. Brackets play a fundamental role in physics where the following are arguably the two most famous brackets:

1. $\mathcal{B}$ is a Riemannian co-metric, denoted $\mathcal{B} \equiv -G^{-1}$, then $X_H = -\nabla H$ is a gradient vector field, and $\dot\gamma = X_H(\gamma)$ defines the Riemannian gradient flow which can be used for minimising H on $\mathcal{F}$ (see §5.2).

2. $\mathcal{B}$ is a Poisson structure, denoted $\mathcal{B} \equiv \Pi$, the associated $\{\cdot,\cdot\}$ is a Poisson bracket, and $X_f \equiv i_{\mathrm{d}f}\Pi = \{f, \cdot\}$ is called a (non-canonical) Hamiltonian vector field (see §4.3.1.5). If the Poisson bracket arises from a symplectic structure, $\Pi = \omega^{-1}$, we will say it is a symplectic bracket.

In the sequel we will commonly consider pairs $(\mathcal{B}, \mu_{\mathcal{F}})$ consisting of a bracket and a smooth measure supported on $\mathcal{F}$, and refer to them as **$\mathcal{B}$-volume manifolds**. We will also regularly use the following two brackets: any vector field Y defines a canonical symmetric bracket $\mathcal{S}$ by

$$\mathcal{S}^\sharp(\alpha) \equiv \alpha(Y)Y, \qquad \{f, h\} = Y(f)Y(h), \tag{4.10}$$

and these symmetric brackets will naturally appear when we discuss diffusions. When we construct Stein discrepancies it will be useful to note that any bracket defines a **dual bracket** $\mathcal{B}^*$ by $\mathcal{B}^{*\sharp} = \mathcal{B}^{\sharp *}$, meaning that $\mathcal{B}^{*\sharp}$ is the adjoint of $\mathcal{B}^\sharp$ with respect to the dual pairing $\left\langle \mathcal{B}^\sharp(\nu), \gamma \right\rangle_* = \left\langle \nu, \mathcal{B}^{*\sharp}(\gamma) \right\rangle_*$, from which it follows that $\mathcal{B}^*(\nu, \gamma) = \mathcal{B}(\gamma, \nu)$.

4.2.2.2 Canonical Transformations

Given a structure, here a bracket, the maps that preserve it, $F^*\mathcal{B} = \mathcal{B}$, are often of interest (later we will refer to such maps as Bracket morphisms). In particular this is the case when F corresponds to the action of a Lie group, since we know that the presence of symmetries is directly related to our ability to efficiently implement a method.

Vector fields Y satisfying $Y\{g, h\} = \{Yg, h\} + \{g, Yh\}$ will be called **canonical vector fields** of the bracket. These form a Lie subalgebra of $\Gamma(T\mathcal{M})$, and are precisely the vector fields satisfying

$$\mathcal{L}_Y \mathcal{B} = 0,$$

or in other words $(\Phi_t^Y)^*\mathcal{B} = \mathcal{B}$, for all t in the domain of Y (by corollary 6.4.2 [251]), as shown below:

Proof. By definition $\{f, g\} = \mathcal{B}(\mathrm{d}f, \mathrm{d}g) = \langle \mathcal{B}, \mathrm{d}f \otimes \mathrm{d}g \rangle$, and since the Lie derivative commutes with contractions and exterior derivatives of functions, $Y\{f, g\} = \mathcal{L}_Y\{f, g\} = \mathcal{L}_Y \langle \mathcal{B}, \mathrm{d}f \otimes \mathrm{d}g \rangle = \langle \mathcal{L}_Y\mathcal{B}, \mathrm{d}f \otimes \mathrm{d}g \rangle + \langle \mathcal{B}, \mathrm{d}\mathcal{L}_Y f \otimes \mathrm{d}g \rangle + \langle \mathcal{B}, \mathrm{d}f \otimes \mathrm{d}\mathcal{L}_Y g \rangle = \langle \mathcal{L}_Y\mathcal{B}, \mathrm{d}f \otimes \mathrm{d}g \rangle + \{Yf, g\} + \{f, Yg\}$. Hence $\mathcal{L}_Y\mathcal{B} = 0$ iff Y is canonical ■

Canonical vector fields are the infinitesimal generators of automorphism of $\mathcal{B}$. They are called Poisson/Symplectic vector fields when $\mathcal{B}$ is Poisson/symplectic, and it turns out the rotationnel of a Poisson structure Π (with respect to any measure) is a canonical vector field as a result of

Jacobi identity [102] (a key result in the construction of the Poisson cohomology). In particular, if $\mathcal{G}$ is a Lie group acting on $\mathcal{F}$, and the bracket is $\mathcal{G}$-invariant, then the infinitesimal generators of the action are canonical vector fields. The existence of a "momentum map" for an action by symplectomorphisms (or Poisson maps) may be understood has simply the requirement that these infinitesimal generators are Hamiltonian vector fields. Let us briefly discuss the relation between $\mathcal{B}$-Hamiltonian and canonical vector fields. For general brackets, $\mathcal{B}$-Hamiltonian vector fields are not necessarily canonical; indeed

$$X_f \text{ is canonical iff } \mathrm{Jac}(f,\cdot,\cdot) \text{ vanishes,} \tag{4.11}$$

where $\mathrm{Jac}(f,g,h) \equiv \{f,\{g,h\}\} - \{\{f,g\},h\} - \{g,\{f,h\}\}$. It follows from Jacobi identity that Hamiltonian vector fields defined by Poisson brackets are canonical. On the other hand, when $\mathcal{B} \equiv G^{-1}$ the canonical vector fields are the Killing vector fields (infinitesimal generators of isometries), but the G^{-1}-Hamiltonian vector fields are the gradient fields, and these do not necessarily generate isometries. In fact, the following derivation shows that in the context of Riemannian brackets, (4.11) recovers the result that a gradient flow ∇f is Killing iff $\mathrm{Hessian}(f) = 0$:

Proof. Recall the Hessian satisfies $H_f(X,Y) = g(\nabla_X \nabla f, Y)$, then using that ∇ is a metric connection and $H_f(X,Y) = H_f(Y,X)$ so $2H_f(X,Y) = g(\nabla_X \nabla f, Y) + g(X, \nabla_Y \nabla f)$

$$\begin{aligned}
\mathrm{Jac}(f,r,h) &= X_f g(\nabla A, \nabla h) - X_h g(\nabla f, \nabla A) - X_A g(\nabla f, \nabla h) \\
&= -2H_f(\nabla A, \nabla h) + H_A(\nabla f, \nabla h) + H_h(\nabla A, \nabla f) - H_A(\nabla f, \nabla h) - H_h(\nabla A, \nabla f) \\
&= -2H_f(\nabla A, \nabla h),
\end{aligned}$$

thus $\mathrm{Jac}(f,\cdot,\cdot) = 0$ iff $\mathrm{Hessian}(f) = 0$. ■

This shows that the Hessian of a function f may be thought as a measure of the failure of the derivation property of the gradient field ∇f. For non-degenerate antisymmetric brackets we can further characterise when a canonical vector field will be locally $\mathcal{B}$-Hamiltonian, as explained in the following subsubsection.

4.2.2.3 Non-degenerate Brackets

We finish this subsection by discussing $\mathcal{B}$-Hamiltonians that will be particularly useful when discussing Stein discrepancies - the Casimirs. As mentioned in §3.2.2, in the case of symplectic brackets, the dynamics associated to H is unique (up to a locally constant function), that is given X_H we can reconstruct $[H]$. However, this is no longer true for general brackets since

their dynamics have a "gauge freedom" determined by the Casimirs: we say f is a **Casimir** when $X_f = 0$, i.e., f is in the kernel of the linear map $\mathcal{B}^\sharp \circ \mathrm{d}$. We will say that $\mathcal{B}$ is **non-degenerate** when the map $\mathcal{B}^\sharp : \Gamma(T^*\mathcal{F}) \to \mathfrak{X}(\mathcal{F})$, $\alpha \mapsto i_\alpha \mathcal{B}$ is invertible at each point. Locally constant functions are always Casimirs, and the converse holds for non-degenerate brackets. A non-degenerate bracket defines a tensor field $\mathcal{B}^{-1}$ on vector fields satisfying

$$\mathcal{B}^{-1}(i_\alpha \mathcal{B}, i_\gamma \mathcal{B}) \equiv \mathcal{B}(\gamma, \alpha), \qquad \Longrightarrow \qquad \mathcal{B}^{-1}(X_f, X_h) = \{h, f\}, \tag{4.12}$$

for any 1-forms α, γ, from which if follows that the $\mathcal{B}$-Hamiltonian vector field associated to f satisfies the **$\mathcal{B}^{-1}$-Hamilton's equation**

$$i_{X_f} \mathcal{B}^{-1} = \mathrm{d}f.$$

For example, the Riemannian bracket $\mathcal{B}$ is invertible since $\mathcal{B}^\sharp$ reduces to the musical isomorphism, and we have $\langle \nabla f, \nabla h \rangle = \{h, f\} = \langle \mathrm{d}f, \mathrm{d}h \rangle^{-1}$, while in the symplectic case $\Pi(\mathrm{d}f, \mathrm{d}h) = \{f, h\} = \omega(X_h, X_f)$. Hence non-degenerate brackets give us a way to think about both Riemannian and Symplectic geometry at the same time. For example, these share a common description of locally $\mathcal{B}$-Hamiltonian vector fields:

Proposition 6. *If $\mathcal{B}$ is non-degenerate, then X is locally $\mathcal{B}$-Hamiltonian iff $\alpha_X \equiv i_X \mathcal{B}^{-1}$ is closed. If $\mathcal{B}^{-1}$ is further antisymmetric, we further have*

$$\mathcal{L}_X \mathcal{B}^{-1} = \mathrm{d}\alpha_X + i_X \mathrm{d}\mathcal{B}^{-1},$$

which shows a canonical vector field X is locally $\mathcal{B}$-Hamiltonian iff $i_X \mathrm{d}\mathcal{B}^{-1} = 0$.

Proof. Given a vector field X and its associated "dual" 1-form $\alpha_X \equiv i_X \mathcal{B}^{-1}$, we have $\mathrm{d}\alpha_X = 0$, iff α_X is locally $\mathrm{d}H$, iff locally $\mathcal{B}^{-1}(X, \cdot) = \mathrm{d}H = \mathcal{B}^{-1}(X_H, \cdot)$, iff $X = X_H$ locally. Note that when $\mathcal{B}^{-1}$ is also antisymmetric then using Cartan's magic formula, we can relate the canonical and locally Hamiltonian vector fields $\mathcal{L}_X \mathcal{B}^{-1} = (\mathrm{d}i_X + i_X \mathrm{d})\mathcal{B}^{-1} = \mathrm{d}\alpha_X + i_X \mathrm{d}\mathcal{B}^{-1}$, so a canonical vector field X will be locally Hamiltonian iff $i_X \mathrm{d}\mathcal{B}^{-1} = 0$. ∎

Hence when $\mathcal{B}$ is Riemannian, $i_X \langle \cdot, \cdot \rangle$ is closed iff X is a locally gradient field, and similarly in the symplectic case $i_X \omega$ is closed iff X is locally Hamiltonian. On the other hand, while all symplectic vector fields are locally Hamiltonian, since $\mathrm{d}\omega = 0$, it is not true that Killing fields are locally gradient, or that Poisson vector fields are locally Hamiltonians.[6] It is also interesting to note that when $\mathcal{A}$ is non-degenerate, $\mathcal{A}^{-1}$ generates a volume form just as in

[6] The first Poisson cohomology describes the quotient between canonical vector fields and Hamiltonian vector fields. In the non-degenerate (i.e., symplectic) case it is isomorphic to the de Rham cohomology (theorem 2.1.4 [102]), which gives another way to interpret the statement that symplectic vector fields are locally Hamiltonian.

Table 4.1: A summary of some standard mechanical systems and their properties.

Bracket	Sym/Anti	Jacobi	Degenerate	Canonical	Hamiltonian vector fields
Riemannian G^{-1}	Sym	No	No	Killing	Gradients
Symplectic ω^{-1}	Anti	Yes	No	Symplectic	Hamiltonian
Poisson Π	Anti	Yes	Yes	Poisson	non-canonical Hamiltonian
$\mathcal{A}$-mechanics	Anti	No	Yes	Canonical	$\mathcal{A}$-Hamiltonian field

the symplectic case. However, unlike for the symplectic case, $\mathcal{A}$-Hamiltonian vector fields do not necessarily preserve this measure, precisely because they are not necessarily canonical. Similarly, the important fact that the commutator $[X, Y]$ of canonical vector fields X, Y is generated by the $\mathcal{A}$-Hamiltonian $\mathcal{A}^{-1}(X, Y)$ requires that $i_X i_Y \mathrm{d}\mathcal{A}^{-1} = 0$ (as can be seen from the proof of proposition 18.3 [93]), and thus holds in general only for symplectic systems. Table 4.1 contains a summary of the properties of standard mechanical systems.

Finally we mention another useful formula which follows immediately from Stokes' theorem (which we will use to derive $\mathcal{B}$-score matching §6.10), and nicely depends on the properties of the bracket: the $\mathcal{B}$-integration by parts formula.

Proposition 7. *For any bracket $\mathcal{B}$, smooth measure μ, 1-form α, and function f, the following $\mathcal{B}$-integration by parts formula holds provided both sides are well-defined and the boundary term vanishes*

$$\int \mathcal{B}(\alpha, \mathrm{d}f)\mu = -\int f \, div_\mu(i_\alpha \mathcal{B})\mu.$$

Thus, using (4.12),

$$\underbrace{\int \mathcal{B}^{-1}(X_f, X)\mu = -\int f \, div_\mu(X)\mu}_{\text{non-degenerate } \mathcal{B}}, \qquad \underbrace{\int \{f, g\}\mu = -\int f \, curl_\mu(\mathcal{A})(g)\mu}_{\mathcal{A}\text{-antisymmetric}}.$$

In the Riemannian and symplectic cases we recover

$$\int \langle \nabla f, X \rangle \, vol = -\int f \nabla \cdot X \, vol, \qquad \int \omega(X_f, X_g)\omega^n = 0.$$

Note the application of the Π-integration by parts (i.e., Π is Poisson) with respect to a Hamiltonian measure μ (defined in §4.3.1.3) generalises (4.34) in [91].[7]

[7] Indeed, as we shall see later the fact that μ is a Hamiltonian measure is equivalent to Π being in the kernel of curl_μ; from proposition 7 we see that whenever the bracket has vanishing curl_μ, for example $\mathcal{A} = \mathrm{curl}_\mu(\mathcal{V})$, then

$$\int \{f, g\}\mu = 0.$$

4.3 The Geometric Mechanics Foundations of HMC

Let us now apply the geometric theory of measures and bracket formalism to sampling. Consider a target distribution P on $\mathcal{M}$ known up to normalisation. While theorem 3 characterises the space of P-preserving dynamics, it is not a priori evident how to construct tractable and efficient target-preserving dynamics. In §3 we have thoroughly analysed HMC, and we have seen that in order to sample from the target distribution $P = e^{-V}\mu_{\mathcal{M}}$, with V known up to a constant, HMC relied on lifting P to a phase space $\mathcal{F}$ and constructing a mechanical system adapted to P with the following properties:

$$\text{we have a } \mathbb{R}\text{-linear map } f \mapsto X_f \text{ depending on } f \text{ (the log-density) up to a constant.} \tag{4.13a}$$

$$X_f(f) = 0 \text{ for any } f. \tag{4.13b}$$

$$\exists \mu_{\mathcal{F}} \text{ s.t., } \mathcal{L}_{X_f}\mu_{\mathcal{F}} = \mu_{\mathcal{F}} \text{ for any } f, \text{ i.e., } \mathfrak{X}_{\mathcal{B}}(\mathcal{F}) \subset \mathrm{Dyn}(\mu_{\mathcal{F}}). \tag{4.13c}$$

$$\text{The "shadow Hamiltonian property": } \forall f, h\ \exists \mathcal{H} \text{ s.t., } \Phi^f_t \circ \Phi^h_t = \Phi^{\mathcal{H}}_t. \tag{4.13d}$$

The set of dynamics $\{X_f\}$ will be said to have **the Shadow Property for splitting integrators** when (4.13b)-(4.13d) hold. With these properties, we can construct an efficient sampling algorithm by first lifting P to a "canonical distribution" $\mu_H \equiv e^{-H}\mu_{\mathcal{F}}$ on $\mathcal{F}$ with H known up to constant, and use (4.13a) to define X_H, which by (4.13b) and (4.13c) preserves μ_H. If the dynamics of X_H is tractable, we can then alternate between "deterministic" steps which follow the trajectories of X_H, but stay within a probability density level set, and heat baths (generated by the disintegration) which shifts between the level sets of e^{-H}. Since X_H is usually intractable, we approximately integrate the dynamics using a splitting scheme by decomposing $H = \sum_i H_i$, so that by $\mathbb{R}$-linearity in (4.13a) we have $X_H = \sum_i X_{H_i}$, and split X_H into X_{H_i}. Then property (4.13d) guarantees the modified equations of the spitting method are still "Hamiltonian", and thus preserve a "volume measure" $\mu_{\mathcal{F}}$ by (4.13c), which ensures no Jacobian factor appear in the accept/reject step. However the splitting integrator no longer preserves μ_H since it does not preserve H. Fortunately property (4.13b) guarantees the Shadow Hamiltonian generating the splitting integrator is exactly preserved by the numerical integrator. Since the splitting method is consistent for X_H, the Shadow Hamiltonian is "close" to H, from which it follows that the energy change during the numerical integration remains "small" even though the numerical trajectory diverges from the correct one – thus ensuring the acceptance rate remains high.

In §4.3.1 we shall analyse the properties (4.13a)-(4.13d), and show that these are all satisfied for Poisson-Hamiltonian pairs $(\Pi, \mu_{\mathcal{F}})$, where Π is a Poisson structure and $\mu_{\mathcal{F}}$ an invariant measure for the associated Poisson mechanics - a generalisation of the symplectic measure

we used in our construction of Hamiltonian Monte Carlo §3.3. In that case, setting H s.t., $\mu_H = e^{-H}\mu_{\mathcal{F}}$, we will show the μ_H-preserving dynamics $\text{curl}_{\mu_H}(\Pi)$ generated by the Poisson structure Π (which is the modular vector field [358]) reduces to the Poisson mechanics X_H associated to the density e^{-H}. The properties of the Poisson stucture, will further ensure that splitting integrators for X_H will be volume-preserving and have good energy-preserving properties as discussed above. In §4.3.2 we will briefly discuss generating functions of $\mathcal{B}$-Hamiltonian morphisms, and re-interpret HMC from the perspective of brackets. Once we have constructed volume-preserving integrators with the shadow property as above, all we need is to compose them with an appropriate "time-reversal symmetry" $S : \mathcal{F} \to \mathcal{F}$ which ensures the deterministic proposal is an involution, and the Metropolis step is well-defined. We will discuss time-reversibility of $\mathcal{B}$-Hamiltonian systems and derive the general bracket MCMC scheme in §4.3.3.

4.3.1 Analysis of the HMC Desiderata

Let us now analyse the properties (4.13). In §4.3.1.1 we will interpret brackets as maps from unnormalised probability densities to dynamical systems depending on the score pointwise, and note the role played by antisymmetry of the bracket to characterise (4.13b) in §4.3.1.2. In §4.3.1.3 we will introduce an important operator, $\phi^{\mathcal{B}}_{\mu_{\mathcal{F}}}$, describing the compatibility between a bracket $\mathcal{B}$ and a reference measure $\mu_{\mathcal{F}}$, and show it reduces to the curl vector field, $\phi^{\mathcal{A}}_{\mu_{\mathcal{F}}} = -\text{curl}_{\mu_{\mathcal{F}}}(\mathcal{A})$, when the bracket is antisymmetric, $\mathcal{B} = \mathcal{A}$, and to the modular vector field of Weinstein [358] when $\mathcal{B}$ is Poisson. In particular we will characterise $\mathcal{B}$-mechanics that have an invariant measure, by proving these must be locally curled, and show that when $\phi^{\mathcal{A}}_{\mu_{\mathcal{F}}}$ vanishes, then X_H preserves the target $e^{-H}\mu_{\mathcal{F}}$. In §4.3.1.4 we will show that the latter is a special instance of a general criteria that characterises whether a "forced system" $X_H + Y$, for some forcing $Y \in \mathfrak{X}(\mathcal{F})$, preserves the target. In §4.3.1.5 we will then discuss the shadow property for splitting method (4.13d) and observe the key role played by Jacobi identity. We will summarise our results in §4.3.1.6 and in particular show that unimodular Poisson manifolds satisfy all the desired properties.

4.3.1.1 From Unnormalised Measures to Dynamics

The first step consists in defining a mapping from target distributions μ (specified up to a positive normalisation) to dynamics. We will parametrise target distributions by their densities with respect to a fixed (but arbitrary) smooth reference measure $\mu_{\mathcal{F}}$, writing $\mu \equiv p_\infty \mu_{\mathcal{F}}$ for some positive density p_∞ known up to normalisation, so that specifying μ up to normalisation corresponds to specifying the equivalence class $[\log p_\infty]$. The brackets defined in §4.2.2 provide

a general mechanism to associate $[\log p_\infty]$ to a dynamics $X_{[\log p_\infty]}$ for which the direction at z is informed by the derivative of the log-density, that is the score:

Proposition 8. *If $\mathcal{F}$ is connected (or restricting to a connected component), given a reference measure $\mu_\mathcal{F}$ there is a bijection between*

1. *Maps from smooth positive densities p_∞ known up to normalisation to dynamics, s.t., $\log p_\infty \mapsto X_{\log p_\infty}$ is $\mathbb{R}$-linear, and the direction of motion depends only on the score pointwise, that is $X_{\log p_\infty}(z)$ only depends on the score $\mathrm{d}\log p_\infty$ through its value evaluation $\mathrm{d}_z \log p_\infty$, and,*

2. *Brackets $\mathcal{B}$ as defined in §4.2.2.*

Proof. Since p_∞ is positive we can identify it with its log-density, and p_∞ is known up to normalisation iff $\log p_\infty$ is known up to constant. Hence, working with log-densities, maps in (1) correspond to $\mathbb{R}$-linear maps $C^\infty(\mathcal{F})/\{\text{constants}\} \to \mathfrak{X}(\mathcal{F})$, $[f] \mapsto X_{[f]}$, s.t., $X_{[f]}(z)$ only depends on f through $\mathrm{d}_z f$. If $\mathcal{F}$ is connected, then there is a bijection between functions defined up to constant and exact 1-forms, namely, $[f] \mapsto \mathrm{d}f$ (for non-connected manifolds this result holds by identifying functions related by locally constant functions, $C^\infty(\mathcal{F})/\{\text{locally constant functions}\}$, 9.3 [93]). Now, given a bracket $\mathcal{B}$ the map $\mathcal{B}^\sharp$ restricted to the vector space of exact 1-forms (over $\mathbb{R}$) provides a map $[f] \mapsto X_{[f]} \equiv \mathcal{B}^\sharp(\mathrm{d}f)$ satisfying (1). Conversely, identifying log-densities defining up to constant with exact 1-forms as above, let $C^\infty(\mathcal{F})/\{\text{constants}\} \to \mathfrak{X}(\mathcal{F}) : [h] \mapsto X_{[h]}$ be a map satisfying (1). Define $\mathcal{B}(\mathrm{d}h, \alpha) \equiv \langle \alpha, X_{[h]} \rangle$, which is $\mathbb{R}$-linear map on exact 1-forms in the first entry and C^∞-linear map on 1-forms in the second. Then for any f (but really $[f]$) define $\tilde{X}_f(x)$ by $\tilde{X}_f(x)(\mathrm{d}_x h) \equiv \mathcal{B}(\mathrm{d}h, \mathrm{d}f)(x) = X_{[h]}(x)(f)$, which is well-defined since $X_{[h]}(x)$ depends on $[h]$ only through $\mathrm{d}_x h$; since $\tilde{X}_f(x)$ is $\mathbb{R}$-linear over exact 1-forms, which form a basis of the cotangent space at x, this determines the vector field $\tilde{X}_f$ (that is tangent vectors are specified by their contraction with exact 1-forms and $\mathbb{R}$-linearity). Then note that $\mathcal{B}(\mathrm{d}h, \mathrm{d}f)(x)$ only depends on $\mathrm{d}_x h$, $\mathrm{d}_x f$, since $\mathcal{B}(\mathrm{d}h, \mathrm{d}f)(x) = \langle \mathrm{d}f, X_{[h]} \rangle (x) = \mathrm{d}_x f(X_{[h]}(x))$ and $\mathcal{B}(\mathrm{d}h, \mathrm{d}f)(x) = \left\langle \mathrm{d}h, \tilde{X}_f \right\rangle (x) = \mathrm{d}_x h(\tilde{X}_f(x))$, and is $\mathbb{R}$-bilinear. Hence $\mathcal{B}$ uniquely determines a $C^\infty(\mathcal{F})$-bilinear map on 1-forms. ∎

This results motivates the use of brackets in the statistical context: whenever the target decomposes into independent components, $p_\infty = \frac{1}{Z}\Pi_j \rho_j$ with Z unknown, the $\mathbb{R}$-linearity in the score implies the dynamical system naturally splits, $X_{\log p_\infty} = \sum_j X_{\log \rho_j}$, and a splitting method may be used to implement $X_{\log p_\infty}$ in practice. Thus writing $\mu \equiv \mu_H \equiv e^{-H}\mu_\mathcal{F}$, the map $\mathcal{B}^\sharp \circ \mathrm{d} : H \mapsto X_H$ defined by a bracket $\mathcal{B}$ satisfies our first desideratum (4.13a).

4.3.1.2 Density-preserving Dynamics

Now that we have achieved desideratum (4.13a), we turn our attention to desideratum (4.13b). The following straightforward argument proves that the latter holds, that is the dynamics X_H associated to the density e^{-H} will remain on the initial probability level set, iff $\mathcal{B}$ is antisymmetric.

Proposition 9. *$X_f(f) = 0$ for all f if and only if $\{\cdot,\cdot\}$ is antisymmetric.*

Proof. $X_f(f) \equiv \{f, f\} = 0$ for any f, if and only if $\{f + h, f + h\} = 0$ for any f and h if and only if $\{f, h\} = -\{h, f\}$. ■

This result explains the central role played by antisymmetric matrix (fields) in sampling methods, and is consistent with our result that antisymmetric brackets naturally arise as local representations of μ-preserving dynamics (theorem 3); we will refer to the dynamics $\mathfrak{X}_{\mathcal{A}}(\mathcal{F})$ generated by an antisymmetric bracket as an **$\mathcal{A}$-mechanics**. Indeed when $\mathcal{B} = G^{-1}$, the Hamiltonian vector field $X_H = \nabla H$ not only does not preserve H, but its integral curves follow the direction of greatest change of H with respect to the Riemannian metric (§5.2). In the following subsection we will further see that only $\mathcal{A}$-mechanics can have an invariant measure.

4.3.1.3 Measure-preserving Flows and Invariant Measures of $\mathcal{B}$-mechanics

Next, we consider (4.13c), the problem of determining the existence of an invariant measure $\mu_{\mathcal{F}}$ for $\mathfrak{X}_{\mathcal{B}}(\mathcal{F})$, i.e., $\mathrm{div}_{\mu_{\mathcal{F}}}(X_f) = 0$ for all $X_f \in \mathfrak{X}_{\mathcal{B}}(\mathcal{F})$. Let $\phi_{\mu_{\mathcal{F}}} : C^\infty(\mathcal{F}) \to C^\infty(\mathcal{F})$ be the $\mathbb{R}$-linear differential operator defined by

$$\phi_{\mu_{\mathcal{F}}} \equiv \mathrm{div}_{\mu_{\mathcal{F}}} \circ \mathcal{B}^\sharp \circ \mathrm{d}, \qquad \phi_{\mu_{\mathcal{F}}}(f) \equiv \mathrm{div}_{\mu_{\mathcal{F}}}(X_f). \tag{4.14}$$

The kernel of $\phi_{\mu_{\mathcal{F}}}$ contains the functions whose $\mathcal{B}$-Hamiltonian vector fields are $\mu_{\mathcal{F}}$-preserving, $\mathcal{B}^\sharp \circ \mathrm{d}\left(\ker(\phi_{\mu_{\mathcal{F}}})\right) \subset \mathrm{Dyn}(\mu_{\mathcal{F}})$. The operator is a vector field iff $\mathcal{B}$ is antisymmetric since

$$\phi_{\mu_{\mathcal{F}}}(fh) = \mathrm{div}_{\mu_{\mathcal{F}}} X_{fh} = \mathrm{div}_{\mu_{\mathcal{F}}}(fX_h + hX_f) = \{h, f\} + f\mathrm{div}_{\mu_{\mathcal{F}}} X_h + \{f, h\} + h\mathrm{div}_{\mu_{\mathcal{F}}} X_f, \tag{4.15}$$

in which case we will call it the **curl vector field** of $(\mathcal{A}, \mu_{\mathcal{F}})$, as it may be written as the $\mu_{\mathcal{F}}$-rotationnel of $\mathcal{A}$ as we show below

$$\phi_{\mu_{\mathcal{F}}} = \mathrm{curl}_{\mu_{\mathcal{F}}}(\mathcal{A}).$$

Proof. If $[\cdot,\cdot]$ denotes the Schouten bracket, X is a multi-vector field and B is a b-vector field,

then [197]

$$[X, B] = (-1)^b \text{curl}_{\mu_\mathcal{F}}(X \wedge B) - \text{curl}_{\mu_\mathcal{F}}(X) \wedge B - (-1)^b X \wedge \text{curl}_{\mu_\mathcal{F}}(B). \tag{4.16}$$

Letting $X \equiv \text{curl}_{\mu_\mathcal{F}}(\mathcal{A})$ and $B \equiv f$ (an arbitrary function), since $[X, f] = i_{\mathrm{d}f} X$ (this holds for any multi-vector field X by (2.75) [102]), we have (using $\text{curl}_{\mu_\mathcal{F}}(f) = 0$ since $\mathrm{d}(f\mu_\mathcal{F}) = 0$)

$$\text{curl}_{\mu_\mathcal{F}}(\mathcal{A})(f) = \text{curl}_{\mu_\mathcal{F}}(f \text{curl}_{\mu_\mathcal{F}}(\mathcal{A})),$$

while letting $X \equiv \mathcal{A}$, $B \equiv f$ and taking the curl of both sides we have

$$\text{curl}_{\mu_\mathcal{F}}([\mathcal{A}, f]) = -\text{curl}_{\mu_\mathcal{F}}(f \text{curl}_{\mu_\mathcal{F}}(\mathcal{A})) = -\text{curl}_{\mu_\mathcal{F}}(\mathcal{A})(f).$$

The result then follows from $\phi_{\mu_\mathcal{F}}(f) = \text{curl}_{\mu_\mathcal{F}}(X_f) = -\text{curl}_{\mu_\mathcal{F}}(i_{\mathrm{d}f}\mathcal{A}) = -\text{curl}_{\mu_\mathcal{F}}([\mathcal{A}, f]) = \text{curl}_{\mu_\mathcal{F}}(\mathcal{A})(f)$, and the fact vector fields are defined by their action on functions. ■

We thus have the following result, which shows the rotationnel of $\mathcal{A}$ with respect to $\mu_\mathcal{F}$ describes whether there are obstructions to $\mu_\mathcal{F}$ being an invariant measure for the $\mathcal{A}$-mechanics.

Corollary 10. *Given an $\mathcal{A}$-mechanics and smooth positive measure $\mu_\mathcal{F}$, the curl vector field $\phi_{\mu_\mathcal{F}}$ is $\mu_\mathcal{F}$-preserving, and vanishes iff $\mu_\mathcal{F}$ is an invariant measure of $\mathfrak{X}_\mathcal{A}(\mathcal{F})$.*

Moreover, just as we characterised measure-preserving dynamics in theorem 3 using "bracket potential" $\mathcal{A}$, we can characterise $\mathcal{B}$-mechanics that preserve a given measure, since once again this question comes back to characterising the kernel of the rotationnel, but this time on 2-vector fields.

Theorem 11. *Let $\mu_\mathcal{F}$ be a smooth positive measure. Then $\mu_\mathcal{F}$ is an invariant measure for the $\mathcal{B}$-mechanics $\mathfrak{X}_\mathcal{B}(\mathcal{F})$ iff $\mathcal{B} = \mathcal{A}$ is locally curled. Moreover, the space of brackets for which $\mu_\mathcal{F}$ is an invariant measure is*

$$\ker\left(curl_{\mu_\mathcal{F}}|_{\mathfrak{X}^2(\mathcal{F})}\right) \cong curl_{\mu_\mathcal{F}}\left(\mathfrak{X}^3(\mathcal{F})\right) \oplus \mu_\mathcal{F}^\sharp\left(H^{n-2}\left(\mathcal{F}\right)\right).$$

Proof. From (4.15) we see that that $\mathcal{B}$ has an invariant measure only if it is antisymmetric. From corollary 10 $\mu_\mathcal{F}$ is an invariant measure iff $\text{curl}_{\mu_\mathcal{F}}(\mathcal{A}) = 0$ iff $\mathrm{d}\mu_\mathcal{F}^\flat(\mathcal{A}) = 0$ iff $\mu_\mathcal{F}^\flat(\mathcal{A}) = \mathrm{d}\gamma$ locally iff $\mathcal{A} = \text{curl}_{\mu_\mathcal{F}}(\mathcal{V})$ locally, for some (local) 3-vector field $\mathcal{V} \equiv \mu_\mathcal{F}^\sharp(\gamma)$. Similarly the result holds globally iff the top-2 de Rham cohomology group is trivial (with a twist in the non-orientable case). ■

Note that using (4.16), if $\mathcal{A} \equiv \text{curl}_{\mu_\mathcal{F}}(\mathcal{V})$ for some $\mathcal{V} \in \Gamma(\Lambda^3 T\mathcal{F})$, we can check with an

analogous proof that in this case

$$X_f = -i_{\mathrm{d}f}\mathcal{A} = -[\mathcal{A}, f] = -[\mathrm{curl}_{\mu_\mathcal{F}}(\mathcal{V}), f] = \mathrm{curl}_{\mu_\mathcal{F}}([\mathcal{V}, f]) = \mathrm{curl}_{\mu_\mathcal{F}}(i_{\mathrm{d}f}\mathcal{V}), \tag{4.17}$$

from which it is clear that any $\mathcal{A}$-Hamiltonian vector field X_H is $\mu_\mathcal{F}$-preserving, and is in fact also $e^{-H}\mu_\mathcal{F}$-preserving due antisymmetry (see lemma 12 below). For a tensor field $\mathcal{R}$ whose first index is contravariant, and a map $\mathcal{K}$ between appropriate bundle sections (for example a differential operator), we can think of the map $[H] \mapsto \mathcal{K}(i_{\mathrm{d}H}(\mathcal{R}))$ as a general mechanism to assign geometric objects to functions known up to constant, such as log-likelihoods in the context of unnormalised statistical models. The choices $\mathcal{K} = \mathrm{Id}$, and $\mathcal{R} = \mathcal{A}$ yield density-preserving dynamics as we saw in §4.3.1.2, while $\mathcal{K} = \mathrm{curl}_{\mu_\mathcal{F}}$, $\mathcal{R} = \mathcal{V}$ generates measure-preserving dynamics. In §6.9 we will further see that the choices $\mathcal{K} = \mathrm{div}_{\mu_\mathcal{F}}$ and $\mathcal{R} = \mathcal{A} + \mathcal{S}$, with $\mathcal{S}$ a symmetric non-negative bracket, allows us to generalise the generator approach for the construction of Stein operators.

We will say that the volume measure $\mu_\mathcal{F}$ is $\mathcal{A}$-**Hamiltonian** if it is an invariant measure of the $\mathcal{A}$-mechanics, or just Hamiltonian when it is a Π-Hamiltonian measure of a Poisson structure Π, in order to be consistent with [91]. Such triples $(\mathcal{F}, \mathcal{A}, \mu_\mathcal{F})$ will be called a **bracket-Hamiltonian manifold**, and bundles $\pi : (\mathcal{F}, \mathcal{A}, \mu_\mathcal{F}) \to \mathcal{M}$ provide the general systems to "implement" idealised/exact-HMC: we define the canonical distribution μ_H from the target measure $P = e^{-V}\mu_\mathcal{M}$ (specified up to a constant) by lifting P using a smooth probability disintegration (as we did for HMC in §3.3)

$$H \equiv -\log \frac{\mathrm{d}(\pi^* P \wedge \zeta)}{\mathrm{d}\mu_\mathcal{F}} = -\log \mu_\mathcal{F}^\sharp \left(e^{-\pi^* V}\pi^*\mu_\mathcal{M} \wedge \zeta\right) = \pi^* V - \log \frac{\mathrm{d}(\pi^*\mu_\mathcal{M} \wedge \zeta)}{\mathrm{d}\mu_\mathcal{F}},$$

and alternate between X_H flows (which preserve $\mu_H = e^{-H}\mu_\mathcal{F}$), and (measure-preserving) heat baths associated to the disintegration ζ. Such systems also posses natural microcanonical distributions on the hypersurfaces of constant probability density, which can be constructed as in §3.2.2.1 by replacing the symplectic measure with $\mu_\mathcal{F}$.

Note that when we pushforward a general canonical distribution μ_H with respect to a diffeomorphism $\mathcal{W} \in \mathrm{Diff}(\mathcal{F})$, we have $\mathcal{W}_\sharp\left(e^{-H}\mu_\mathcal{F}\right) = e^{-H\circ\mathcal{W}^{-1}}|\mathcal{J}(\mu_\mathcal{F}, \mathcal{W}^{-1})|\mu_\mathcal{F}$, so

$$\mathcal{W}_\sharp^{-1}\mu_H = \mu_{\mathcal{W}^* H - \log|\mathcal{J}(\mu_\mathcal{F}, \mathcal{W})|}, \tag{4.18}$$

which yields the general formula to transport samples between canonical distributions. In particular if $\mathcal{W}$ corresponds to the flow of a $\mathcal{A}$-Hamiltonian vector field X_W, $\mathcal{W} = \Phi_t^W$, and $\mu_\mathcal{F}$ is $\mathcal{A}$-Hamiltonian, then the transformation simplifies to $(\Phi_t^W)_\sharp\mu_H = \mu_{(\Phi_{-t}^W)^* H}$. Next we show that if $\mu_\mathcal{F}$ is not $\mathcal{A}$-Hamiltonian, one can still add a "forcing" vector field to the Hamiltonian

system to ensure the target measure is preserved. This "forcing" is precisely the curl vector field defined above.

4.3.1.4 Measure-preserving Bracket Dynamics with Forcing

One can consider more complicated non-linear associations of unnormalised densities to dynamics, by for example adding a forcing vector field Y, so $H \mapsto X_H + Y$. This may be useful when we do not have a $\mathcal{B}$-Hamiltonian measure $\mu_{\mathcal{F}}$ for $\mathcal{B}$. We then fix an arbitrary one, say $\mu_{\mathcal{F}}$, and set $\mu_H = p_\infty \mu_{\mathcal{F}} = e^{-H}\mu_{\mathcal{F}}$. What are the conditions on the forcing Y to ensure $X_H + Y$ preserves μ_H? The answer relies on the following lemma:

Lemma 12. *For any bracket $\mathcal{B}$,*

$$\phi^{\mathcal{B}}_{e^{-H}\mu_{\mathcal{F}}} = \phi^{\mathcal{B}}_{\mu_{\mathcal{F}}} - X^{\mathcal{B}^*}_H,$$

so if $\mathcal{B} \equiv \mathcal{A}$ is antisymmetric the curl vector fields of $e^{-H}\mu_{\mathcal{F}}$ and $\mu_{\mathcal{F}}$ are related by $\phi_{e^{-H}\mu_{\mathcal{F}}} = \phi_{\mu_{\mathcal{F}}} + X_H$.

Proof. Recall $\phi_{e^{-H}\mu_{\mathcal{F}}}(f) \equiv \mathrm{div}_{e^{-H}\mu_{\mathcal{F}}} X_f$. Now

$$\mathcal{L}_{X_f} e^{-H}\mu_{\mathcal{F}} = X_f(e^{-H})\mu_{\mathcal{F}} + e^{-H}\mathcal{L}_{X_f}\mu_{\mathcal{F}} = \left(\mathrm{div}_{\mu_{\mathcal{F}}}(X_f) - X_f(H)\right) e^{-H}\mu_{\mathcal{F}},$$

Hence, $\mathrm{div}_{e^{-H}\mu_{\mathcal{F}}}(X^{\mathcal{B}}_f) = \mathrm{div}_{\mu_{\mathcal{F}}}(X^{\mathcal{B}}_f) - X^{\mathcal{B}}_f(H) = \mathrm{div}_{\mu_{\mathcal{F}}}(X^{\mathcal{B}}_f) - X^{\mathcal{B}^*}_H(f)$, so $\phi^{\mathcal{B}}_{e^{-H}\mu_{\mathcal{F}}} = \phi^{\mathcal{B}}_{\mu_{\mathcal{F}}} - X^{\mathcal{B}^*}_H$. Moreover if $\mathcal{B} = \mathcal{A}$, $\mathrm{div}_{e^{-H}\mu_{\mathcal{F}}}(X_f) = \mathrm{div}_{\mu_{\mathcal{F}}}(X_f) + X_H(f)$, from which the result follows. ■

The previous lemma shows that the curl vector field of the target measure μ naturally decomposes into the curl vector field of the reference measure and the Hamiltonian dynamics generated by the probability density, whenever μ is expanded as a canonical distribution $\mu \equiv \mu_H \equiv e^{-H}\mu_{\mathcal{F}}$. Moreover it shows that if $\mu_{\mathcal{F}}$ is $\mathcal{A}$-Hamiltonian, then $e^{-h}\mu_{\mathcal{F}}$ is $\mathcal{A}$-Hamiltonian iff h is a Casimir, and thus a locally constant function in the case of non-degenerate brackets. Combined with corollary 10, it also shows that the forced dynamics $X_H + \phi_{\mu_{\mathcal{F}}}$ is μ_H-preserving, and thus that X_H is μ_H-preserving when $\mu_{\mathcal{F}}$ is $\mathcal{A}$-Hamiltonian. The next proposition characterises measure-preserving forced $\mathcal{B}$-dynamics, in terms of the distortion of volume:

Proposition 13. *The forced $\mathcal{B}$-Hamiltonian dynamics $X_H + Y$ preserves $\mu_H = p_\infty \mu_{\mathcal{F}}$ iff*

$$div_{\mu_{\mathcal{F}}}(X_{p_\infty} - p_\infty Y) = 0, \text{ or equivalently} \quad \phi_{\mu_{\mathcal{F}}}(p_\infty) = div_{\mu_{\mathcal{F}}}(p_\infty Y).$$

In particular if $\mathcal{B}$ is antisymmetric, $X_H + Y$ is target-preserving iff $div_{\mu_{\mathcal{F}}}(p_\infty(\phi_{\mu_{\mathcal{F}}} - Y)) = 0$.

Proof. Computing the divergence gives us

$$\begin{aligned}\mathcal{L}_{X_H+Y}\mu_H &= e^{-H}\mathcal{L}_{X_H}\mu_{\mathcal{F}} + X_H(e^{-H})\mu_{\mathcal{F}} + Y(e^{-H})\mu_{\mathcal{F}} + e^{-H}\mathcal{L}_Y\mu_{\mathcal{F}} \\ &= \left(\mathrm{div}(X_H) - X_H(H) - Y(H) + \mathrm{div}Y\right)\mu_H,\end{aligned}$$

so the flow of $X_H + Y$ preserves μ_H iff

$$\mathrm{div}(X_H) - X_H(H) - Y(H) + \mathrm{div}Y = 0,$$

iff

$$\mathrm{div}(X_{p_\infty} - p_\infty Y) = 0.$$

If $\mathcal{B}$ is antisymmetric, we further have (using (4.27))

$$\mathrm{div}_{\mu_{\mathcal{F}}}(X_{p_\infty}) = \phi_{\mu_{\mathcal{F}}}(p_\infty) = \phi_{\mu_{\mathcal{F}}}(p_\infty) + p_\infty \mathrm{div}_{\mu_{\mathcal{F}}}(\phi_{\mu_{\mathcal{F}}}) = \mathrm{div}_{\mu_{\mathcal{F}}}(p_\infty \phi_{\mu_{\mathcal{F}}}),$$

so the condition now reads

$$\mathrm{div}_{\mu_{\mathcal{F}}}(p_\infty(\phi_{\mu_{\mathcal{F}}} - Y)) = 0,$$

of which $Y \equiv \phi_{\mu_{\mathcal{F}}}$ is a solution.[8] ■

So indeed by forcing X_H with the curl vector field of the reference measure, we generate dynamics that preserve μ_H. More generally, setting $Y = \phi_{\mu_{\mathcal{F}}} + Z$, we can further add a vector field Z s.t., Z/p_∞ is not globally $\mu_{\mathcal{F}}$-curled when the first homology group of P is non-trivial. We will see an example of this when we analyse measure-preserving diffusions on compact manifolds (see §4.4.3), using the isomorphism between the de Rham cohomology and the harmonic forms.

4.3.1.5 Modified Equations for Splitting Methods

In practice a bracket-Hamiltonian system will not be sufficient to provide efficient sampling as the flow of X_H will be intractable. In that case we typically resort to a splitting method, and will thus further need the Shadow property to hold. Hence we discuss conditions under which splitting methods have a shadow vector Hamiltonian, i.e., desideratum (4.13d) holds. By the Baker-Campbell-Hausdorff formula (see §3.5.4.2), the (formal) existence of a shadow Hamiltonian is guaranteed by requiring that $[X_f, X_g] = X_l$, i.e., $\mathfrak{X}_{\mathcal{B}}(\mathcal{F})$ forms a Lie subalgebra of vector fields. Let us define the Jacobiator by

$$\mathrm{Jac}(f, g, h) \equiv \{f, \{g, h\}\} - \{\{f, g\}, h\} - \{g, \{f, h\}\}, \tag{4.19}$$

[8] We could have also relied on corollary 10 and lemma 12, $\mathcal{L}_{X_H+\phi_{\mu_{\mathcal{F}}}}\mu_H = \mathcal{L}_{\phi_{\mu_H}}\mu_H = 0$.

which measures the lack of "derivation property" of bracket fields:

$$X_f\{g,h\} = \{X_f g, h\} + \{g, X_f h\} + \mathrm{Jac}(f,g,h).$$

Then[9],

$$[X_f, X_g]h = X_f\{g,h\} - X_g\{f,h\} = -\mathrm{Jac}(f,g,h) + X_{\{f,g\}}h.$$

Hence, for $\mathfrak{X}_{\mathcal{B}}(\mathcal{F})$ to form a Lie subalgebra of $\mathfrak{X}(\mathcal{F})$, the RHS in the above expression must be of the form $X_l h$, which is equivalent to requiring that $\mathrm{Jac}(f,g,\cdot) \in \mathfrak{X}_{\mathcal{B}}(\mathcal{F})$ for all f,g. This is satisfied for instance when $\mathrm{Jac}(\cdot,\cdot,\cdot) = 0$ and $l = \{f,g\}$, that is, $(C^\infty(\mathcal{F}), \{\cdot,\cdot\})$ forms a left Leibniz algebra [238]. In particular, if $\mathcal{B}$ is also antisymmetric the Jacobiator (4.19) reduces to the usual one from Poisson mechanics, and $\mathrm{Jac} = 0$ is equivalent to demanding that $\mathcal{B}$ is a Poisson bi-vector field, $\mathcal{B} = \Pi$. When the bracket $\mathcal{B} \equiv \Pi$ is Poisson, $(\mathcal{F}, \Pi)$ is referred to as a Poisson manifold, and the corresponding bracket dynamics $X_H \equiv i_{\mathrm{d}H}\Pi$ is known as a *Poisson mechanics*. This encompasses a wider class of dynamics than the "canonical" Hamiltonian systems that are defined by a symplectic structure (i.e., $\Pi \equiv \omega^{-1}$). We say that a Poisson system is "non-canonical" if it does not arise from a symplectic structure, and such systems commonly arise in the context of reduction by symmetry. Some examples of non-canonical systems in mechanics include the Euler's equations for rigid body motion, or Euler's equations for ideal fluids [271].

4.3.1.6 Brackets with the Shadow Property for Splitting Integrators

As mentioned at the beginning of the section, requiring $\mathfrak{X}_{\mathcal{B}}$ to have the shadow property for splitting integrators is equivalent to demanding that (i) the shadow vector field is $\mathcal{B}$-Hamiltonian, i.e., the $\mathcal{B}$-Hamiltonian vector fields form a Lie subalgebra, (ii) that the shadow Hamiltonian is preserved by the flow, i.e., the bracket is antisymmetric, (iii) and that a $\mathcal{B}$-Hamiltonian measure exists (in order to avoid the appearance of expensive Jacobian factors in the Metropolis ratio). We saw the first two conditions are satisfied when $\mathcal{B}$ is a Poisson bivector field. A triple $(\mathcal{F}, \Pi, \mu_{\mathcal{F}})$ where Π is a Poisson bivector field, and $\mu_{\mathcal{F}}$ is a Hamiltonian measure will be called a **Poisson Hamiltonian system**. These are the natural mechanical systems to implement sampling methods with HMC-like properties. Naturally, given a manifold $\mathcal{M}$, the cotangent bundle $(T^*\mathcal{M}, \omega^{-1}, \omega^n)$ is a Poisson-Hamiltonian manifold (where ω is the canonical symplectic structure), and when $\mathcal{M}$ is Riemannian, we can translate it to velocity space which yields the Riemann Newton space $(T\mathcal{M}, \omega_\flat^{-1}, \omega_\flat^n)$.
For Poisson manifolds, the existence of a Hamiltonian measure is called "unimodularity" and

[9] which also shows the Jacobiator is a derivation over its last entry

is described for example in [358, 101, 233, 58, 111, 195]. In that case, the curl vector field ϕ_μ is called the **modular vector field**. It defines a 1-cocycle in the Lichnerowicz complex, and its cohomology class in the first Poisson cohomology is called the modular class (the existence of this cohomology relies critically on the Jacobi identity). The Hamiltonian vector fields are coboundaries and lemma 12 implies that the curl vector fields of all measures belong to the same cohomology class; in other words, the modular class does not depend on μ, and thus characterises whether an invariant measure exists: $(\mathcal{F}, \Pi)$ admits a smooth invariant measure iff the modular class is trivial.

In summary, if $(\mathcal{F}, \Pi)$ is a unimodular Poisson bundle over $\mathcal{M}$, let $\mu_\mathcal{F}$ be a Hamiltonian measure and lift P to a target measure $\mu = e^{-H}\mu_\mathcal{F}$. Then the μ-preserving dynamics ϕ_μ reduces to X_H by lemma 12, that is the reference measure $\mu_\mathcal{F}$ does not appear. By Jacobi identity any splitting integrator Ψ for X_H will have a modified equation given by a shadow Hamiltonian vector field $X_{\mathcal{H}_\Psi}$, which preserves the shadow Hamiltonian $\mathcal{H}_\Psi$ by antisymmetry, $X_{\mathcal{H}_\Psi}(\mathcal{H}_\Psi) = 0$, and $\mu_\mathcal{F}$. Hence, since $\mathrm{div}_\mu(X_{\mathcal{H}_\Psi}) = \mathrm{div}_{\mu_\mathcal{F}}(X_{\mathcal{H}_\Psi}) - \{\mathcal{H}_\Psi, H\} = \mathrm{div}_{\mu_\mathcal{F}}(X_{\mathcal{H}_\Psi}) - \mathrm{d}H(X_{\mathcal{H}_\Psi})$, the Jacobian factor $\mathcal{J}(\mu, \Phi_{\delta t}^{X_{\mathcal{H}_\Psi}})$ of the splitting integrator with respect to the target μ

$$\mathcal{J}(\mu, \Phi_{\delta t}^{X_{\mathcal{H}_\Psi}})(z) = e^{\int_0^{\delta t} \mathrm{div}_\mu(X_\mathcal{H})(\Phi_s^{X_{\mathcal{H}_\Psi}}(z))\mathrm{d}s} = e^{-H\left(\Phi_{\delta t}^{X_{\mathcal{H}_\Psi}}(z)\right)+H(z)}\mathcal{J}(\mu_\mathcal{F}, \Phi_{\delta t}^{X_{\mathcal{H}_\Psi}})(z), \tag{4.20}$$

is just a function of the energy difference along the flow of the shadow Hamiltonian since $\mathcal{J}(\mu_\mathcal{F}, \Phi_{\delta t}^{X_{\mathcal{H}_\Psi}}) = 1$ from the fact that $\mu_\mathcal{F}$ is Hamiltonian.

4.3.2 Shadow of Symplectic Integrators and HMC

We saw in the context of HMC that any consistent symplectic map has a shadow Hamiltonian, not just the symplectic maps associated to splitting methods (recall §3.5.4.1). Therein we mentioned that the construction of the shadow Hamiltonian relies on the observation that any symplectic map has an associated generating function, from which one can then construct a differential equation, the Hamilton-Jacobi PDE, which describes the time-evolution that a generating function must satisfy in order to define the flow of some Hamiltonian vector field (see also §A.5.6.4). Given the Hamilton-Jacobi PDE we can then obtain an asymptotic expansion (in the time-step) for the Shadow Hamiltonian of any symplectic integrator consistent with a Hamiltonian system. It is thus natural to wonder whether more general $\mathcal{B}$-morphisms are also generated by a function.

Proposition 14. *Let $\pi_i : \mathcal{F}\times\mathcal{F} \to \mathcal{F}$, $A \equiv \pi_2^*\mathcal{B} - \pi_1^*\mathcal{B}$, $F : \mathcal{F} \to \mathcal{F}$, and $\iota_F : \mathrm{graph}(F) \to \mathcal{F}\times\mathcal{F}$ inclusion. Then F canonical (a $\mathcal{B}$-morphism) iff $\iota_F^* A = 0$. For Poisson brackets this is the statement that a map is Poisson iff its graph is a coisotropic submanifold of $(\mathcal{F}, \mathcal{B}) \times (\mathcal{F}, -\mathcal{B})$.*

Proof. Recall that $F : \mathcal{F} \to \mathcal{F}$ is canonical iff $F^*\mathcal{B} = \mathcal{B}$. Then

$$\pi_1 \circ \iota_F = \pi_1|_{\mathrm{graph}(F)} \text{ and } \pi_2 \circ \iota_F = F \circ \pi_1|_{\mathrm{graph}(F)}.$$

Thus

$$\iota_F^* A = \pi_1|^*_{\mathrm{graph}(F)} F^*\mathcal{B} - \pi_1|^*_{\mathrm{graph}(F)} \mathcal{B} = \pi_1|^*_{\mathrm{graph}(F)} (F^*\mathcal{B} - \mathcal{B})$$

so F canonical iff $\iota_F^* A = 0$. ■

Thus the characterisation of canonical maps in terms of their graph, which is the first step in the construction of Hamilton-Jacobi PDE, is common to all brackets. However, it does not seem possible to derive a generating function in general. Indeed the symplectic case relies on the fact that $\mathcal{B}$ is non-degenerate antisymmetric, so that we can reproduce the above derivation with the two form $\mathcal{B}^{-1}$, to express $\iota_F^* A = 0$ in the form $\iota_F^* (\pi_2^* \mathcal{B}^{-1} - \pi_1^* \mathcal{B}^{-1}) = 0$, and then use the fact that $\mathcal{B}^{-1}$ is locally exact (that is closed), $\mathcal{B}^{-1} = \mathrm{d}\alpha$, to show that this implies that locally $\mathrm{d}\iota_F^* (\pi_2^* \alpha - \pi_1^* \alpha) = 0$, which further implies the local existence of a generating function $W : \mathrm{graph}(F) \to \mathbb{R}$, s.t., $\iota_F^* (\pi_2^* \alpha - \pi_1^* \alpha) = \mathrm{d}W$. In fact, while there are "generating functions" for contact systems and a "Lie–Poisson Hamilton–Jacobi", both rely on the standard Hamilton-Jacobi PDE of symplectic transformations. Specifically, in the Lie–Poisson case the Lie-Poisson Hamilton Jacobi theorem is constructed by assuming the generating function is invariant under the Lie group action, and thus symplectically reduces to the co-Lie algebra [371], giving rise to a left-trivialised version of Hamilton-Jacobi theorem. In the case of contact transformations, the derivation of the PDE relies on the symplectisation of exact contact manifolds [117, 60].

We have now developed a clear understanding of the role played by the properties of Hamiltonian mechanics for the purpose of HMC. For example:

- the fact Hamiltonian mechanics is a $\mathcal{B}$-dynamics ensures we do not need to know the normalisation of the target probability density $\frac{1}{Z} e^{-H}$, that X_H is informed by the score $\mathrm{d}H$, and that we can implement X_H via a splitting integrator;
- the antisymmetry of ω ensures that the probability density e^{-H} is preserved by X_H;
- the Jacobi property ensures splitting integrators have a $\mathcal{B}$-Hamiltonian shadow Hamiltonian vector field (whose Hamiltonian is preserved by antisymmetry);
- the non-degeneracy and antisymmetry of ω induces a natural reference measure ω^n, which is further invariant because Jacobi identity ensures that Hamiltonian vector fields are canonical (and unique up to locally constant functions); the existence of an invariant measure and antisymmetry implies the existence of a microcanonical distribution on the probability density level sets;
- combining non-degeneracy with antisymmetry and Jacobi identity finally ensures that symplectic maps have a generating function, guaranteeing the existence of a shadow

Hamiltonian for all (consistent) symplectic integrators.

4.3.3 Time-Reversibility and Bracket MCMC

Once we have a deterministic proposal associated to a $\mathcal{B}$-Hamiltonian system, we need to turn it into an involution in order to obtain a well-defined Metropolis accept-reject step. In HMC such an involution is typically given by the momentum flip, which is geometrically an antisymplectic involution (recall §3.4.1). We now consider in full generality such transformations of the brackets. We say a C^1-diffeomorphism $S : \mathcal{F} \to \mathcal{F}$ is a **Bracket (anti)-morphism** if[10]

$$S^*\{f, g\} = \pm\{S^*f, S^*g\}, \quad \text{or equivalently} \quad S^*\mathcal{B} = \pm\mathcal{B}.$$

Then just as for Poisson maps, we have[11]

$$\partial S \circ X_{S^*f} = \pm X_f \circ S, \quad \text{and thus} \quad S \circ \Phi_t^{S^*f} = \Phi_{\pm t}^f \circ S.$$

Proof.

$$\begin{aligned}(\partial S X_{S^*f})[h](S(z)) &= X_{S^*f}(z)[S^*h] = X_{S^*f}[S^*h](z) = \{S^*f, S^*h\}(z) \\ &= \pm S^*\{f, h\}(z) = \pm\{f, h\}(S(z)) = \pm X_f[h](S(z)).\end{aligned}$$

■

Suppose now that S is an antimorphism, and we have a reversible Hamiltonian, $S^*H = H$. Then

$$S \circ \Phi_t^H = \Phi_{-t}^H \circ S, \quad \text{or} \quad \Phi_{-t}^H = S \circ \Phi_t^H \circ S^{-1}.$$

This is known as a time-discrete symmetry of the $\mathcal{B}$-dynamics [201]. When S is involutive, $S^{-1} = S$, as in §3.4.1 any integrator with such property will be called **time-reversible**, and we will say it is **time-symmetric** if $\Psi_{\delta t}^{-1} = \Psi_{-\delta t}$ (note that given any integrator $\Psi_{\delta t}$, the integrator $\Psi_{\frac{1}{2}\delta t} \circ \Psi_{-\frac{1}{2}\delta t}^{-1}$ is time-symmetric). If both property holds, the deterministic proposal $S \circ \Psi_{\delta t}$ is an involution

$$(S \circ \Psi_{\delta t})^{-1} = \Psi_{\delta t}^{-1} \circ S^{-1} = \Psi_{-\delta t} \circ S = S \circ \Psi_{\delta t}.$$

More generally $S \circ \Psi_{\delta t}$ is involutive iff the integrator $\Psi_{\delta t}$ satisfies $\Psi_{\delta t}^{-1} = S \circ \Psi_{\delta t} \circ S$ for some bijection $S : \mathcal{F} \to \mathcal{F}$, and we will refer to this property as S**-reversibility**. In particular if S

[10] Technically, we should denote $S^*\mathcal{B}$ by $S_*\mathcal{B}$ since it really corresponds to the pushforward of the tensor $\mathcal{B}$.
[11] i.e., X_{S^*f} and X_f are S-related. Recall $Z, Y \in \mathfrak{X}(\mathcal{F})$ are S-related iff $Y(S^*f) = Z(f) \circ S \ \forall f$

is an involutive antimorphism, and $H = \sum_i H_i$, we can define the splitting method

$$\Psi_{\delta t} \equiv \Phi_{\delta t}^{H_1} \circ \cdots \circ \Phi_{\delta t}^{H_\ell}.$$

If all the Hamiltonians are reversible then $\Psi_{\delta t}$ is time-reversible, and it is time-symmetric provided the sequence $(H_1, \ldots, H_\ell)$ is palindromic. Thus $\Psi_{\delta t}$ is S-reversible.

Given a volume manifold $(\mathcal{M},P)$, it is straightforward to generalise [114] to arbitrary manifolds, and obtain the following **compressible bracket MCMC** algorithm[12]:

Proposition 15. *Given a target $P = e^{-V}\mu_{\mathcal{M}}$ on $\mathcal{M}$, let $\pi : (\mathcal{F}, \mu_{\mathcal{F}}) \to (\mathcal{M}, \mu_{\mathcal{M}})$ be a fibre bundle and fix a probability disintegration $\zeta \in \Omega^{\dim \mathcal{F} - \dim \mathcal{M}}(\mathcal{F})$. Set $\mu_H \equiv e^{-H}\mu_{\mathcal{F}}$ where $H \equiv \pi^* V - \log \frac{\mathrm{d}(\pi^*\mu_{\mathcal{M}} \wedge \zeta)}{\mathrm{d}\mu_{\mathcal{F}}}$. Suppose $\Psi_{\delta t}$ is a S-reversible diffeomorphism, where S is a μ_H-preserving bijection. The following algorithm generates a μ_H-preserving Markov chain. Given $z^\ell \equiv (q^\ell, p^\ell)$*

1. *$p_*^\ell \leftarrow p^\ell$ where p_*^ℓ is sampled from $\zeta|_{\pi^{-1}(q^\ell)}$*
2. *$z_* \leftarrow \Psi_{\delta t}(q^\ell, p_*^\ell)$*
3. *set $z^{\ell+1} \leftarrow z_*$ with probability $\min\left(1, e^{-H(z_*)+H(q^\ell,p_*^\ell)}|\mathcal{J}(\mu_{\mathcal{F}}, \Psi_{\delta t})|\right)$, else $z^{\ell+1} \leftarrow S(q^\ell, p_*^\ell)$,*

where $\mathcal{J}(\mu_{\mathcal{F}}, \Psi_{\delta t})$ is the Jacobian of $\Psi_{\delta t}$ with respect to the positive one-density $\mu_{\mathcal{F}}$. Then the projections $\pi(z^\ell)$ are realisations from a P-invariant Markov chain.

Proof. We adapt the proof in appendix A [114] to general manifolds and reference measures. The transition kernel of the last two steps above is $\mathbb{P}(z \to \mathrm{d}z^*) \equiv \min(1, r(z, z^*))\, \delta_{\Psi_{\delta t}(z)}(\mathrm{d}z^*) + (1 - \min(1, r(z, \Psi_{\delta t}(z)))\, \delta_{S(z)}(\mathrm{d}z^*)$, where $r(z, z^*) = e^{-H(z_*)+H(z)}|J(\mu_{\mathcal{F}}, \Psi_{\delta t})|(z)$, and we need to show $\int \mathbb{P}(z \to \mathrm{d}z^*)\mu_H(\mathrm{d}z) = \mu_H(\mathrm{d}z^*)$. One term of $\int \mathbb{P}(z \to \mathrm{d}z^*)\mu_H(\mathrm{d}z)$ is

$$\int \delta_{S(z)}(\mathrm{d}z^*)\mu_H(\mathrm{d}z) = \int \delta_z(\mathrm{d}z^*)(S^{-1})^*\mu_H(\mathrm{d}z) = \int \delta_z(\mathrm{d}z^*)\mu_H(\mathrm{d}z) = \mu_H(\mathrm{d}z^*).$$

Another term is (using $\Psi_{\delta t} \circ S \circ \Psi_{\delta t} = S$, and $S \circ \Psi_{\delta t}$ is an involution)

$$\begin{aligned}
&\int \min(1, r(z, z^*))\, \delta_{\Psi_{\delta t}(z)}(\mathrm{d}z^*)\mu_H(\mathrm{d}z) = \int \min(1, r(z, \Psi_{\delta t}(z)))\, \delta_{\Psi_{\delta t}(z)}(\mathrm{d}z^*)\mu_H(\mathrm{d}z) \\
&= \int \min\left(1, e^{-H(\Psi_{\delta t}(z))+H(z)}|\mathcal{J}(\mu_{\mathcal{F}}, \Psi_{\delta t})|\right) \delta_{\Psi_{\delta t}(z)}(\mathrm{d}z^*)\mu_H(\mathrm{d}z) \\
&= \int \min\left(e^{-H(z)}, e^{-H(\Psi_{\delta t}(z))}|\mathcal{J}(\mu_{\mathcal{F}}, \Psi_{\delta t})|\right) \delta_{\Psi_{\delta t}(z)}(\mathrm{d}z^*)\mu_{\mathcal{F}}(\mathrm{d}z) \\
&= \int \min\left(e^{-H\circ S\circ\Psi_{\delta t}(z)}, e^{-S^*H(z))}|J(\mu_{\mathcal{F}}, \Psi_{\delta t}) \circ S \circ \Psi_{\delta t}|\right) \delta_{S(z)}(\mathrm{d}z^*)(S \circ \Psi_{\delta t})^*\mu_{\mathcal{F}}(\mathrm{d}z)
\end{aligned}$$

[12] The following result does not require S to be involutive, but simply that it is S-reversible in the more general sense $\Psi_{\delta t}^{-1} \circ S = S \circ \Psi_{\delta t}$.

$$= \int \min\left(e^{-H\circ S\circ\Psi_{\delta t}(z)}, e^{-S^*H(z))}|J(\mu_{\mathcal{F}},\Psi_{\delta t})\circ S\circ\Psi_{\delta t}|\right)\delta_{S(z)}(\mathrm{d}z^*)|\mathcal{J}(\mu_{\mathcal{F}},S\circ\Psi_{\delta t})|(z)\mu_{\mathcal{F}}(\mathrm{d}z)$$
$$= \int \min\left(e^{-H\circ S\circ\Psi_{\delta t}(z)}|\mathcal{J}(\mu_{\mathcal{F}},S\circ\Psi_{\delta t})|(z), e^{-S^*H(z))}|J(\mu_{\mathcal{F}},S)(z)|\right)\delta_{S(z)}(\mathrm{d}z^*)\mu_{\mathcal{F}}(\mathrm{d}z)$$
$$= \int \min\left(e^{-H\circ S\circ\Psi_{\delta t}(z)}|\mathcal{J}(\mu_{\mathcal{F}},S\circ\Psi_{\delta t})|(z), e^{-H(z)}\right)\delta_{S(z)}(\mathrm{d}z^*)\mu_{\mathcal{F}}(\mathrm{d}z)$$
$$= \int \min\left(e^{-H\circ S\circ\Psi_{\delta t}(z)}e^{-H+S^*H}\circ\Psi_{\delta t}(z)|\mathcal{J}(\mu_{\mathcal{F}},\Psi_{\delta t})|(z), e^{-H(z)}\right)\delta_{S(z)}(\mathrm{d}z^*)\mu_{\mathcal{F}}(\mathrm{d}z)$$
$$= \int \min\left(e^{-H\circ\Psi_{\delta t}(z)}|\mathcal{J}(\mu_{\mathcal{F}},\Psi_{\delta t})|(z), e^{-H(z)}\right)\delta_{S(z)}(\mathrm{d}z^*)\mu_{\mathcal{F}}(\mathrm{d}z),$$

which exactly cancels with the last term; here we have used the properties of the Jacobian (proposition 7.5.14 [3]), and that $\mathcal{J}(\mu_{\mathcal{F}},S\circ\Psi_{\delta t}) = \mathcal{J}(\mu_{\mathcal{F}},S)\circ\Psi_{\delta t}\mathcal{J}(\mu_{\mathcal{F}},\Psi_{\delta t}) = e^{-H+S^*H}\circ\Psi_{\delta t}\mathcal{J}(\mu_{\mathcal{F}},\Psi_{\delta t})$, since $S^*\mu_H = \mu_H$ implies $S^*\mu_{\mathcal{F}} = e^{-H+S^*H}\mu_{\mathcal{F}}$. ■

In particular if we choose a bracket $\mathcal{B}$ on $\mathcal{F}$, and let S be an involutive $\mu_{\mathcal{F}}$-preserving antimorphism, if H is reversible then S preserves μ_H, and the splitting method obtained by splitting H into a palindromic sequence of reversible Hamiltonians is S-reversible, so the above proposition may be used. Moreover if $\mu_{\mathcal{F}}$ is $\mathcal{B}$-Hamiltonian then $\mathcal{J}(\mu_{\mathcal{F}},\Psi_{\delta t}) = 1$, and finally if $(\mathcal{F},\mathcal{B},\mu_{\mathcal{F}})$ is a Poisson-Hamiltonian manifold the algorithm will have the energy-stabilising Shadow property.

4.4 Measure-Preserving Diffusions

We now generalise the forced dynamical systems generated through brackets by adding noise to consider diffusions processes associated to forced $\mathcal{B}$-Hamiltonian dynamics, s.t., the diffusion is measure-preserving whenever the forced $\mathcal{B}$-Hamiltonian dynamics is - this will be the content of §4.4.1. Inspired by our previous results we will then specialise to the class of $\mathcal{A}$-diffusions in §4.4.2, whose deterministic part is given by the curl vector field of the target, and thus only depend on the given target measure. We will see that these are locally complete: any measure-preserving diffusion may be written as an $\mathcal{A}$-diffusion locally. Locally $\mathcal{A}$-diffusions play in the context of measure-preserving diffusions the role that locally curl vector fields played in the context of measure-preserving dynamics, and locally curl brackets in the context of measure-preserving $\mathcal{B}$-mechanics. In §4.4.3 we will then consider the important case when $\mathcal{F}$ is compact, from which we can build harmonic representative of the Fokker-Planck current to obtain global completeness from the Hodge decomposition. In §4.4.4 we will specialise to the Euclidean space and show that $\mathcal{A}$-diffusions recover the complete recipe of diffusions proposed

in [242] - thus showing that the proof of completeness is a simple application of Poincaré lemma once the geometry of measures is taken into account. Finally in §4.4.5 we will consider a symmetric/antisymmetric decomposition of $\mathcal{A}$-diffusion and its generator, as well as discuss transformations under which the generator is reversible - a generalisation of the analogous result for Langevin processes in the case of momentum flip transformation.

4.4.1 Bracket diffusion with forcing

Now that we have developed a geometric understanding of target-preserving dynamics, we can consider the problem of how to add noise. We define the Stratonovich diffusion on $\mathcal{F}$ by considering the process $B \equiv (t, W_t) : \Omega \to \mathbb{R}^{N+1} \equiv U$ where W_t is an N-dimensional Brownian motion, a vector bundle homomorphism $L \in \mathrm{Hom}(U, T\mathcal{F}) = \Gamma(\mathcal{F}, U^* \otimes T\mathcal{F})$, and setting

$$\mathrm{d}Z_t \equiv L(Z_t) \circ \mathrm{d}B = Y_0(Z_t)\mathrm{d}t + Y_i(Z_t) \circ \mathrm{d}W_t^i$$

for vector fields $Y_j \in \Gamma(T\mathcal{F})$ [279]. We consider Stratonovich SDEs since we do not want to depend on a connection, and because the Stratonovich differentials transform naturally under transformations: if $F : \mathcal{F} \to \mathcal{N}$ is a smooth map between manifolds, then (see (35) [279])

$$\mathrm{d}(F \circ Z_t) = \partial F(\mathrm{d}Z_t).$$

Given a target measure μ, expressed as the Gibbs measure of some function $H : \mathcal{F} \to \mathbb{R}$ and inverse temperature $\beta > 0$

$$\mu_H \equiv p_\infty \mu_{\mathcal{F}} = \frac{1}{Z} e^{-\beta H} \mu_{\mathcal{F}}, \quad Z = \int_{\mathcal{F}} e^{-\beta H} \mu_{\mathcal{F}} < \infty, \tag{4.21}$$

we will show that for any bracket $\mathcal{B}$ and vector fields $Y_1, \ldots, Y_N$ on $\mathcal{F}$, the diffusion

$$\mathrm{d}Z_t = \Big(X_H - \frac{\beta}{2} \sum_{i=1}^N Y_i(H) Y_i\Big)\mathrm{d}t + \Big(\tfrac{1}{2} \sum_{i=1}^N \mathrm{div}_{\mu_{\mathcal{F}}}(Y_i) Y_i + Y\Big)\mathrm{d}t + \sum_{i=1}^N Y_i \circ \mathrm{d}W_t^i, \tag{4.22}$$

preserves the Gibbs measure (4.21) provided the vector field Y is chosen such that

$$\mathrm{div}_{\mu_{\mathcal{F}}}(X_{p_\infty} - \beta p_\infty Y) = 0, \tag{4.23}$$

which controls the volume distortion along the vector field X_{p_∞} generated by the target density. By proposition 13 it follows that (4.22) is μ_H-preserving iff the forced $\mathcal{B}$-Hamiltonian system $X_H + Y$ is. The noise terms thus balance themselves, and have a clear interpretation: $Y_i(H)Y_i$ and $\mathrm{div}(Y_i)Y_i$ describe respectively the rate of change of H and the volume deformation along

Y_i. Hereafter, we will refer to the general diffusion process (4.22) as a **bracket diffusion** (written $\mathcal{B}$-diffusion) with forcing Y. To prove the result, we use the following auxiliary result.

Lemma 16. *Given (smooth) vector fields $X, Y_1, \ldots, Y_N$ on a manifold $\mathcal{F}$, consider the Stratonovich SDE*

$$\mathrm{d}Z_t = X(Z_t)\mathrm{d}t + \sum_{i=1}^{N} Y_i(Z_t) \circ \mathrm{d}W_t^i \,. \tag{4.24}$$

Then its generator is given by

$$\mathcal{L}f = Xf + \tfrac{1}{2}\sum_{i=1}^{N} Y_i Y_i f. \tag{4.25}$$

Furthermore, if $\partial\mathcal{F} = \varnothing$ or if we restrict to functions that vanish on the boundary, then the Fokker-Planck operator (defined as the formal adjoint of $\mathcal{L}$ in $L^2(\mathcal{F}, \mu_{\mathcal{F}})$) is

$$\mathcal{L}^* f = div_{\mu_{\mathcal{F}}}\left(-fX + \tfrac{1}{2}\sum_{i=1}^{N} div_{\mu_{\mathcal{F}}}(fY_i)Y_i\right) . \tag{4.26}$$

We will call $-fX + \frac{1}{2}\sum_{i=1}^{N} div_{\mu_{\mathcal{F}}}(fY_i)Y_i$ the $(\mu_{\mathcal{F}}, f)$-Fokker-Planck current of the diffusion.

Proof. The formula for the generator (4.25) is well-known, but we here include a proof, which is a slight generalisation of the Lie group setting in [91], as it will be useful later. Recall first that the generator of the Stratonovich SDE $\mathrm{d}Z_t = b(Z_t)\mathrm{d}t + \sigma(Z_t) \circ \mathrm{d}W_t$ on $\mathbb{R}^n$ is (proposition 3.4 [294])

$$\mathcal{L}f = b^j\partial_j f + \tfrac{1}{2}\sigma_i^k\sigma_i^j\partial_k\partial_j f + \tfrac{1}{2}\sigma_i^k\partial_k\sigma_i^j\partial_j f.$$

In local coordinates the generator of the drift component is $X^i\partial_i f$ which is just Xf. Then consider the noise part $\mathrm{d}Z_t = Y_i \circ \mathrm{d}W_t^i$. If we rewrite this in the form $\mathrm{d}Z_t = \sigma \circ \mathrm{d}W_t$, we have in local coordinates $\sigma_i^j = Y_i^j$. Locally the generator of the diffusion part is $\mathcal{L}f = \frac{1}{2}\sigma_i^k\sigma_i^j\partial_k\partial_j f + \frac{1}{2}\sigma_i^k\partial_k\sigma_i^j\partial_j f$, where $\frac{1}{2}\sigma_i^k\partial_k\sigma_i^j\partial_j$ is the Ito-Stratonovich correction while $\frac{1}{2}\sigma_i^k\sigma_i^j\partial_k\partial_j f$ is the usual generator term associated to the Ito diffusion. For a vector field Y, we have $YY(f) = Y^i\partial_i(Y^j\partial_j f) = Y^iY^j\partial_i\partial_j f + Y^i\partial_i Y^j\partial_j f$. Hence using $\sigma_i^j = Y_i^j$, we find $\frac{1}{2}Y_iY_i f = \frac{1}{2}\sigma_i^k\sigma_i^j\partial_k\partial_j f + \frac{1}{2}\sigma_i^k\partial_k\sigma_i^j\partial_j f$, and it follows that $\mathcal{L}f = \frac{1}{2}Y_iY_i(f)$.

To derive the Fokker-Planck operator we will use the following product rule

$$X(f) = \mathrm{div}(fX) - f\mathrm{div}(X)\,, \tag{4.27}$$

which is well-known when div is the Riemannian divergence, and follows by a slight generalisation of pg 910 [91] (who assumed a Hamiltonian volume form):

Proof of (4.27). Since $\mathcal{L}_{fX}w = f\mathcal{L}_X w + \mathrm{d}f \wedge i_X w$ [13] and $i_X(\mathrm{d}f \wedge \alpha) = (i_X \mathrm{d}f)\wedge\alpha - \mathrm{d}f \wedge i_X\alpha$:

$$\begin{aligned} X(f)\mu_{\mathcal{F}} &= (i_X \mathrm{d}f)\mu_{\mathcal{F}} = (i_X \mathrm{d}f)\mu_{\mathcal{F}} - i_X(\mathrm{d}f \wedge \mu_{\mathcal{F}}) = \\ &= \mathrm{d}f \wedge i_X\mu_{\mathcal{F}} = \mathrm{d}f \wedge i_X\mu_{\mathcal{F}} + f\mathcal{L}_X\mu_{\mathcal{F}} - f\mathcal{L}_X\mu_{\mathcal{F}} \\ &= \mathcal{L}_{fX}\mu_{\mathcal{F}} - f\mathcal{L}_X\mu_{\mathcal{F}} = \big(\mathrm{div}(fX) - f\mathrm{div}(X)\big)\mu_{\mathcal{F}}. \end{aligned}$$

■

Proof of (4.26). Let f, g be arbitrary compactly supported smooth functions. We consider first the drift part. Vector fields satisfy Leibniz rule

$$\int gX(f)\mu_{\mathcal{F}} = \int X(gf)\mu_{\mathcal{F}} - \int fX(g)\mu_{\mathcal{F}},$$

and note that $\int_{\mathcal{M}} \mathrm{div}(gX)\mu_{\mathcal{F}} = \int_{\mathcal{M}} \mathrm{d}i_{gX}\mu_{\mathcal{F}} = \int_{\partial\mathcal{M}} i_{gX}\mu_{\mathcal{F}} = \int_{\partial\mathcal{M}} g i_X\mu_{\mathcal{F}}$ which vanishes since $\partial\mathcal{M} = \varnothing$ (in fact it is sufficient that $g|_{\partial M} = 0$). Hence

$$\begin{aligned} \int (\mathcal{L}f)g\mu_{\mathcal{F}} = \int gX(f)\mu_{\mathcal{F}} &= -\int gf\mathrm{div}(X)\mu_{\mathcal{F}} - \int fX(g)\mu_{\mathcal{F}} \\ &= \int f\big(-g\mathrm{div}(X) - X(g)\big)\mu_{\mathcal{F}} \\ &= \int f\big(-\mathrm{div}(gX)\big)\mu_{\mathcal{F}} = \int f\mathcal{L}^* g\mu_{\mathcal{F}}. \end{aligned}$$

Now let us compute the adjoint of the diffusion component. First note that

$$\begin{aligned} Y_k(gY_k f) &= \mathrm{div}\big(gY_k fY_k\big) - gY_k f\mathrm{div}(Y_k) = \\ &= \mathrm{div}\big(gY_k fY_k\big) - \mathrm{div}\big(fg\mathrm{div}(Y_k)Y_k\big) + f\mathrm{div}\big(g\mathrm{div}(Y_k)Y_k\big) \end{aligned}$$

and

$$Y_i gY_i f = \mathrm{div}(fY_i gY_i) - f\mathrm{div}(Y_i gY_i).$$

It follows that if f, g vanish on the boundary

$$\begin{aligned} \tfrac{1}{2}\int_{\mathcal{M}} gY_iY_i f\mu_{\mathcal{F}} &= \tfrac{1}{2}\int \Big(Y_i\big(gY_i f\big) - Y_i gY_i f\Big)\mu_{\mathcal{F}} \\ &= \tfrac{1}{2}\int \Big(f\mathrm{div}\big(g\mathrm{div}(Y_i)Y_i\big) + f\mathrm{div}(Y_i gY_i)\Big)\mu_{\mathcal{F}} \end{aligned}$$

[13] This follows from Cartan's magic formula, $\mathcal{L}_X = i_X\mathrm{d} + \mathrm{d}i_X$, since $\mathcal{L}_{fX}\alpha = i_{fX}\mathrm{d}\alpha + \mathrm{d}i_{fX}\alpha = fi_X\mathrm{d}\alpha + \mathrm{d}f \wedge i_X\alpha + f\mathrm{d}i_X\alpha = f\mathcal{L}_X\alpha + \mathrm{d}f \wedge i_X\alpha$.

$$= \int f \tfrac{1}{2}\mathrm{div}\big(\mathrm{div}(gY_i)Y_i\big)\mu_{\mathcal{F}} = \int f \mathcal{L}^* g \mu_{\mathcal{F}}.$$

So the adjoints of X and Y_iY_i in $L^2(\mathcal{F}, \mu_{\mathcal{F}})$ are $g \mapsto -\mathrm{div}_{\mu_{\mathcal{F}}}(gX)$ and $g \mapsto \mathrm{div}_{\mu_{\mathcal{F}}}\big(\mathrm{div}_{\mu_{\mathcal{F}}}(gY_i)Y_i\big)$ respectively. ■

■

We now prove the claim that μ_H is a solution of the stationary Fokker–Planck equation, i.e., $\mathcal{L}^* p_\infty = 0$ (see 2 [294]), iff (4.23) holds.

Theorem 17. *Let $\mu_{\mathcal{F}}$ be a smooth positive measure on a manifold $\mathcal{F}$ (without boundary). The Gibbs measure μ_H in equation (4.21) is a stationary distribution of the bracket diffusion (4.22) if and only if $X_H + Y$ is a μ_H-preserving forced dynamics (proposition 13). In addition,*

- ***(i)*** *if $(\mathcal{F}, \mathcal{A}, \mu_{\mathcal{F}})$ is a bracket-Hamiltonian manifold, so $\mu_{\mathcal{F}}$ is Hamiltonian, (4.23) is satisfied in the unforced setting $Y \equiv 0$;*
- ***(ii)*** *more generally for any $\mathcal{A}$-mechanics and $\mu_{\mathcal{F}}$, we can choose $Y \equiv \phi_{\mu_{\mathcal{F}}} = curl_{\mu_{\mathcal{F}}}(\mathcal{A})$. In particular if $\mathcal{F}$ is pseudo-Riemannian, and $\mu_{\mathcal{F}} \equiv vol$ is the Riemannian measure, we can set $Y \equiv -\nabla \cdot \mathcal{A} \equiv -\mathrm{Tr}\nabla\mathcal{A}$.*

Proof. Let $p(x) = \frac{1}{Z}e^{-\beta x}$ for $x \in \mathbb{R}$, so $p_\infty = p \circ H : \mathcal{F} \to \mathbb{R}$. Then using the Fokker-Planck operator defined in (4.26), we have

$$\begin{aligned}
\mathcal{L}^* p_\infty &= \mathrm{div}\left(-p_\infty\Big(X_H - \frac{\beta}{2}Y_i(H)Y_i + \tfrac{1}{2}\mathrm{div}(Y_i)Y_i + Y\Big) + \tfrac{1}{2}\mathrm{div}\big(p_\infty Y_i\big)Y_i \right) \\
&= \mathrm{div}\left(\frac{1}{\beta}X_{p_\infty} - p_\infty Y - \frac{1}{2}Y_i(p_\infty)Y_i - \tfrac{1}{2}p_\infty \mathrm{div}(Y_i)Y_i + \tfrac{1}{2}\mathrm{div}\big(p_\infty Y_i\big)Y_i \right) \\
&= \mathrm{div}\left(\frac{1}{\beta}X_{p_\infty} - p_\infty Y - \frac{1}{2}Y_i(p_\infty)Y_i - \tfrac{1}{2}p_\infty \mathrm{div}(Y_i)Y_i + \tfrac{1}{2}\mathrm{div}\big(p_\infty Y_i\big)Y_i \right) \\
&= \mathrm{div}\left(\frac{1}{\beta}X_{p_\infty} - p_\infty Y - \tfrac{1}{2}\mathrm{div}\big(p_\infty Y_i\big)Y_i + \tfrac{1}{2}\mathrm{div}\big(p_\infty Y_i\big)Y_i \right) \\
&= \mathrm{div}\Big(\frac{1}{\beta}X_{p_\infty} - p_\infty Y\Big),
\end{aligned}$$

thus $\mathcal{L}^* p_\infty = 0$ iff Y satisfies $\mathrm{div}\big(\frac{1}{\beta}X_{p_\infty} - p_\infty Y\big) = 0$.

*Proof of **(i)***: If $(\mathcal{F}, \mathcal{B}, \mu_{\mathcal{F}})$ is a $\mathcal{B}$-Hamiltonian manifold, then $\mathcal{L}_{X_f}\mu_{\mathcal{F}} = 0$ so $\mathrm{div}_{\mu_{\mathcal{F}}}(X_f) = 0$ and we may take $Y = 0$.

*Proof of **(ii)***: If $\mathcal{B}$ is antisymmetric, the result follows by 13. If $(\mathcal{F}, g)$ is pseudo-Riemannian, recall that $\nabla \cdot \mathcal{A} = \mathrm{Tr}\nabla\mathcal{A}$, where Tr is the trace and ∇ the covariant derivative. By using the fact that $S_{ab}\mathcal{B}^{ab} = 0$ for any symmetric tensor S_{ab} and anti-symmetric tensor $\mathcal{B}^{ab}$, we get

$$\begin{aligned}\nabla \cdot \big(p_\infty \nabla \cdot \mathcal{A}\big) &= \frac{1}{\sqrt{|G|}}\partial_b\Big(p\partial_a(\sqrt{|G|}\mathcal{B}^{ab})\Big) = \frac{1}{\sqrt{|G|}}(\partial_b p)\Big(\partial_a(\sqrt{|G|}\mathcal{B}^{ab})\Big) \\ &= \frac{1}{\sqrt{|G|}}\partial_a\Big(\sqrt{|G|}\mathcal{B}^{ab}(\partial_b p)\Big) = -\frac{1}{\sqrt{|G|}}\partial_a\Big(\sqrt{|G|}X_p^a\Big) = -\nabla \cdot X_{p_\infty}\,,\end{aligned}$$

which shows that $Y \equiv -\nabla \cdot \mathcal{A}$ satisfies (4.23). ■

Remark 18 (**Ergodicity**). For the bracket diffusion to be ergodic, it suffices as usual that the noise fields satisfy the Hörmander condition, that is the Lie algebra generated by $\{Y_i, [X, Y_i] : i = 1, \ldots, N\}$ span the tangent spaces, though this is a slightly stronger requirement than necessary when the vector fields are non-analytic (the condition is allowed to fail on appropriate hypersurfaces) [33, 18, 48, 167, 32, 74]. On the other hand, finding the rate of convergence to the stationary distribution is a challenging issue.

4.4.2 Completeness of the $\mathcal{A}$-Diffusions

An $\mathcal{A}$-diffusion forced using the curl vector field $\phi_{\mu_\mathcal{F}} = -\mathrm{curl}_{\mu_\mathcal{F}}\mathcal{A}$ will be simply referred to as an **$\mathcal{A}$-diffusion**

$$\mathrm{d}Z_t = \Big(X_H - \frac{\beta}{2}\sum_{i=1}^N Y_i(H)Y_i\Big)\mathrm{d}t + \Big(\tfrac{1}{2}\sum_{i=1}^N \mathrm{div}_{\mu_\mathcal{F}}(Y_i)Y_i + \phi_{\mu_\mathcal{F}}\Big)\mathrm{d}t + \sum_{i=1}^N Y_i \circ \mathrm{d}W_t^i. \tag{4.28}$$

These diffusions are of fundamental importance in machine learning and statistics since, as the next theorem shows, the class of $\mathcal{A}$-diffusions is locally complete and thus provides the starting point for the construction of measure-preserving diffusions adapted to a particular context.

Theorem 19. *A smooth positive measure μ is a solution of the stationary Fokker–Planck equation of the diffusion*

$$\mathrm{d}Z_t = X(Z_t)\mathrm{d}t + \sum_{i=1}^N Y_i(Z_t) \circ \mathrm{d}W_t^i$$

iff the drift takes the form $X = curl_\mu\mathcal{A} + \frac{1}{2}div_\mu(Y_i)Y_i$ on a neighbourhood of any point for some antisymmetric bracket $\mathcal{A}$. Hence the class of $\mathcal{A}$-diffusions

$$\mathrm{d}Z_t = curl_\mu(\mathcal{A})\,\mathrm{d}t + \tfrac{1}{2}div_\mu(Y_i)Y_i\mathrm{d}t + Y_i \circ \mathrm{d}W_t^i, \tag{4.29}$$

is locally complete. The $\mathcal{A}$-diffusions are globally complete, that is there always exists an $\mathcal{A}$ s.t.,

any measure-preserving diffusion takes the form (4.29) *globally, iff the* $(\dim(\mathcal{F})-1)$*-de Rham cohomology is trivial. Else we need to include the term representing the first homology group of* μ*.*

Proof. Consider the Fokker-Planck operator (4.26) with $\mu = \mu_{\mathcal{F}}$ (so $[p_\infty] = 1$) applied to $f \equiv p_\infty = 1$.[14] Then $\mathcal{L}^*1$ vanishes iff $\operatorname{div}_\mu\left(-X + \frac{1}{2}\sum_{i=1}^N \operatorname{div}_\mu(Y_i)Y_i\right) = 0$ that is the Fokker-Planck current is a locally curled vector field by theorem 3, $-X+\frac{1}{2}\sum_{i=1}^N \operatorname{div}_\mu(Y_i)Y_i \in \operatorname{Curl}_{loc}(\mu)$. Hence $X = \operatorname{curl}_\mu(\mathcal{A}) + \frac{1}{2}\sum_{i=1}^N \operatorname{div}_\mu(Y_i)Y_i$ for some $\mathcal{A}$ on a neighbourhood of any point, and the result follows. ■

The result provides not only the generalisation of the complete recipe derived in [242], it also allows us to re-interpret the derivation and place it into its correct geometric setting. Indeed on $\mathcal{F} = \mathbb{R}^n$ the de Rham cohomology is trivial, so theorem 3 implies the result that if Y is divergence-free (with respect to the Lebesgue measure), $\nabla \cdot Y = 0$, then there exists an antisymmetric matrix $\mathcal{A}$ for which $Y = \nabla \cdot \mathcal{A}$ [194, 260]. Indeed $\mathcal{A}$ is simply the "vector potential" of the "electromagnetic field" Y, which shows that the completness proof of [242] is simply the construction of a vector potential for the Fokker-Planck current. It follows that we can drop the integrability assumption in theorem 2 therein: it is a consequence of the specific choice of vector potential constructed in their proof - see for example [260] for an alternative construction.

4.4.3 $\mathcal{A}$-Diffusions on Riemannian Manifolds

In some cases the de Rham cohomology is non-trivial but it is still possible to obtain a global expression for the drift of any μ-preserving diffusion. This is possible when $\mathcal{F}$ is given a Riemannian metric and the Hodge decomposition of differential forms holds. In that case, as we next prove, the additional term which enables us to achieve global completeness is essentially an harmonic representative of the Fokker-Planck current.

Theorem 20. *Let* $\mathcal{F}$ *be a Riemannian manifold on which the Hodge decomposition of* $\dim(\mathcal{F})-1$*-forms holds (e.g., a compact oriented manifold). Let* $\mu_{\mathcal{F}} \equiv vol$ *and* $\nabla\cdot$ *denote respectively the Riemannian volume form and divergence. Any* e^{-H} *vol-preserving diffusion has the form*

$$\mathrm{d}Z_t \equiv \left(X_H + \tfrac{1}{2}\left(\nabla \cdot Y_i - Y_i(H)\right)Y_i - \nabla\cdot\mathcal{A} + e^H \sharp \star^{-1}\zeta\right)\mathrm{d}t + Y_i \circ \mathrm{d}W_t^i, \tag{4.30}$$

[14] Note that when we work directly with the target measure μ rather than a target in density form $p_\infty\mu_{\mathcal{F}}$, it is perhaps more natural to consider the adjoint of the generator $\mathcal{L}$ with respect to the pairing between compactly supported smooth functions and one-densities defined by integration $\langle f, \mu\rangle \equiv \int f\mu$, instead of (4.26), in which case $\mathcal{L}^*P = -\mathcal{L}_X P + \frac{1}{2}\mathcal{L}_{Y_i}\mathcal{L}_{Y_i}P$.

where $\mathcal{A}$ is antisymmetric and ζ harmonic (it satisfies "Maxwell's equations" $\mathrm{d}\zeta = 0$, $\mathrm{d}\star\zeta = 0$).

Proof. The Fokker-Planck operator (4.26) may be written as

$$\mathcal{L}^* g = \delta \left(-gX + \tfrac{1}{2}\nabla \cdot (gY_i)Y_i \right)^\flat . \tag{4.31}$$

It then follows by the Hodge decomposition that $\mathcal{L}^* p_\infty = 0$ iff $\delta \left(-p_\infty X + \frac{1}{2}\nabla \cdot (p_\infty Y_i)Y_i \right)^\flat = 0$ iff $\mathrm{d} \star \left(-p_\infty X + \frac{1}{2}\nabla \cdot (p_\infty Y_i)Y_i \right)^\flat = 0$ iff

$$\star \left(-p_\infty X + \tfrac{1}{2}\nabla \cdot (p_\infty Y_i)Y_i \right)^\flat = \mathrm{d}\alpha - \gamma$$

where γ is Harmonic, that is co-exact and closed, and vanishes whenever the $\dim(\mathcal{F}) - 1$ de Rham cohomology is trivial. Thus $\left(-p_\infty X + \frac{1}{2}\nabla \cdot (p_\infty Y_i)Y_i \right)^\flat = \delta\varepsilon - \star^{-1}\gamma$ where $\varepsilon \equiv \star^{-1}\alpha$ is a 2-form. Hence

$$-p_\infty X + \tfrac{1}{2}\nabla \cdot (p_\infty Y_i)Y_i = \sharp \delta\varepsilon + \sharp \star^{-1} \gamma = \nabla \cdot (\mathcal{A} p_\infty) - \sharp \star^{-1} \gamma$$

for an appropriate antisymmetric bracket $\mathcal{A}$. Now $\nabla \cdot (\mathcal{A} p_\infty) = p_\infty \nabla \cdot \mathcal{A} - p_\infty X_H$,[15] from which it follows that the drift vector field must take the form

$$X = X_H + \tfrac{1}{2}\nabla \cdot (Y_i)Y_i - \tfrac{1}{2} Y_i(H) Y_i - \nabla \cdot \mathcal{A} + e^H \sharp \star^{-1} \gamma.$$

■

A similar result can be found in [176], although the latter assumes the diffusion is also non-degenerate in order to obtain a Riemannian metric from the noise process (a Riemannian Brownian motion), which is used to turn the Fokker–Planck current into a 1-form that is analysed through its de Rham–Hodge–Kodaira decomposition. The situation is clarified when the assumptions of compactness and non-degeneracy are treated separately, and we will discuss the case of non-degenerate diffusions below. The above theorem shows that no assumption on the noise is necessary, though it assumes that we can express our target measure in terms of the Riemannian measure, which might be inconvenient in practice. Topologically, the additional presence of the harmonic term may be understood from the fact that on compact oriented manifolds, the de Rham cohomology groups are isomorphic to the spaces of harmonic forms. It follows that any P-preserving dynamics has the form $\mathrm{curl}_P(\mathcal{A}) + P^\sharp(\zeta)$ for an antisymmetric bracket $\mathcal{A}$ and non-zero harmonic $(n-1)$-form ζ.

As mentioned above, another context in which a Riemannian metric appears is in the presence

[15] Indeed $\mathrm{curl}_\mu(p_\infty \mathcal{A})[f] = -\mathrm{div}_\mu(p_\infty X_f) = -X_f(p_\infty) - p_\infty \mathrm{div}_\mu X_f = +p_\infty X_f(H) - p_\infty \mathrm{div}_\mu X_f = -p_\infty X_H(f) - p_\infty \mathrm{div}_\mu X_f$, so $\mathrm{curl}_\mu(p_\infty \mathcal{A}) = -p_\infty X_H + p_\infty \mathrm{curl}_\mu \mathcal{A}$

of non-degenerate noise, wherein the generator can be expressed in the form $\mathcal{L} = X + \frac{1}{2}\Delta$, and the noise process generated by the Laplace–Beltrami operator $\frac{1}{2}\Delta$ is the Riemannian Brownian motion.

Theorem 21. *Let vol, ∇ and $\nabla\cdot$ be the Riemannian measure, gradient and divergence respectively. Any $P \propto p_\infty$vol-preserving diffusion generated by $\mathcal{L} = X + \frac{1}{2}\Delta$ takes for some $\mathcal{A} \in \mathfrak{X}^2(\mathcal{M})$, the form (up to topological obstructions)*

$$\mathrm{d}Z_t = \underbrace{-X^{\mathcal{A}}_{\log p_\infty}(Z_t)\mathrm{d}t - \nabla\cdot\mathcal{A}(Z_t)\mathrm{d}t}_{\textit{irreversible drift}} + \underbrace{\overbrace{\tfrac{1}{2}\nabla\log p_\infty(Z_t)\mathrm{d}t}^{\textit{Riemannian gradient flow}} + \overbrace{\mathrm{d}B_t}^{\textit{Riemannian Brownian motion}}}_{\textit{reversible Riemannian overdamped Langevin system}}. \tag{4.32}$$

In the reversible case $\mathcal{A} = 0$, the diffusion (4.32) gives us precisely the *Riemannian overdamped Langevin equation*, used to construct the Riemann Metropolis-adjusted Langevin algorithm (MALA) when the Riemannian metric is obtained from an information divergence [131, 234]. It is also well-known that the presence of the irreversible component ($\mathcal{A} \neq 0$) can accelerate convergence to the target distribution, as demonstrated in [172, 103, 305], with a detailed analysis of the optimal drift for constant $\mathcal{A}$ provided in [213]. Alternatively, reversible overdamped Langevin systems ($\mathcal{A} = 0$) with appropriate choices of Riemannian metric G^{-1} resulting from a Stratonovich perturbation can also lead to accelerated convergence relative to the standard overdamped Langevin process, see [1].

4.4.4 Bracket Diffusion on Flat Euclidean Space

Consider the flat space $\mathcal{F} = \mathbb{R}^n$ with $\mu_{\mathcal{F}}$ the Lebesgue measure, then setting $Y = -\nabla\cdot\mathcal{A} = -\partial_j\mathcal{A}^{ji}\partial_i$ for any anti-symmetric $\mathcal{A}$ recovers the formulation of diffusions preserving the Gibbs distribution that was given in [242]:

Corollary 22. *Let $\mathcal{F} = \mathbb{R}^n$ with flat metric, $\mu_{\mathcal{F}} = \mathrm{d}x$, $\sigma_{ij} \equiv Y^i_j$, $D \equiv \frac{1}{2}\sigma\sigma^T$ and $Q_{ij} \equiv \mathcal{A}^{ij}$. Then (4.22) reduces to the Itô diffusion considered in the complete recipe for SGMCMC [242]*[16]

$$\mathrm{d}Z_t = -\big(Q + D\big)\nabla H\mathrm{d}t + \nabla\cdot\big(Q + D\big)\mathrm{d}t + \sqrt{2D}\mathrm{d}W_t\,, \tag{4.33}$$

Proof. In that case, the term $Y_i(H)Y_i = Y^j_i Y^k_i \partial_j H\partial_k$ represents the "symmetric semi-definite part" of the dynamics since in flat space this is $(\sigma\sigma^T)_{jk}\partial_j H\partial_k$ where $\sigma_{ij} = \sigma^i_j = Y^i_j$, while the term $\mathrm{div}(Y_i)Y_i = \sigma_{ki}\partial_j\sigma_{ji}\partial_k$ together with the Itô-to-Statonovich correction yields the divergence of the diffusion matrix $D \equiv \frac{1}{2}\sigma\sigma^T$. Using the convention $(\nabla\cdot Q)_i \equiv \partial_j Q_{ij}$ in [242]

[16] Here we use the convention $(\nabla\cdot Q)_i \equiv \partial_j Q_{ij}$ as in [242].

$$
\begin{aligned}
\mathrm{d}Z_t &= \left(\mathcal{A}^{ji}\partial_j H\partial_i - \frac{\beta}{2}(\sigma\sigma^T)_{jk}\partial_j H\partial_k + \tfrac{1}{2}\sigma_{ki}\partial_j\sigma_{ji}\partial_k - \partial_j\mathcal{A}^{ji}\partial_i\right)\mathrm{d}t \\
&\qquad + \sigma_{ij}\partial_i \circ \mathrm{d}W_t^j \\
&= \left(\mathcal{A}^{ji}\partial_j H\partial_i - \frac{\beta}{2}(\sigma\sigma^T)_{jk}\partial_j H\partial_k + \tfrac{1}{2}\sigma_{ki}\partial_j\sigma_{ji}\partial_k + \tfrac{1}{2}\sigma_{ki}\partial_k\sigma_{ji}\partial_j + \partial_j\mathcal{A}^{ij}\partial_i\right)\mathrm{d}t \\
&\qquad + \sigma_{ij}\partial_i \mathrm{d}W_t^j \\
&\qquad = \left(-Q\nabla H - \beta D\nabla H + \nabla\cdot D + \nabla\cdot Q\right)\mathrm{d}t + \sqrt{2D}\mathrm{d}W_t.
\end{aligned}
$$

■

4.4.5 Dissipative Form of $\mathcal{A}$-diffusions and Hamiltonian Noise

Let us now discuss the symmetric/antisymmetric decomposition of $\mathcal{A}$-diffusions. Recall that any vector field Y defines a canonical symmetric bracket by $\{f,g\} \equiv Y(f)Y(g)$, see (4.10). Let $\mathcal{S}$ denote the symmetric bracket obtained by summing the canonical brackets of the noise fields, i.e., $\mathcal{S}^\sharp(\alpha) = \alpha(Y_i)Y_i$. In particular viewing $\mathcal{S}^\sharp$ as a vector field-valued 1-form on $\Omega^1(\mathcal{F})$ we may think of the μ-divergence as acting only on the vector field component of $\mathcal{S}$, which yields a 1-form on $\Omega^1(\mathcal{F})$, and then use the canonical isomorphism $T\mathcal{F} \cong T^{**}\mathcal{F}$ between tangent spaces and their double dual to obtain a vector field. This leads us to the definition $\mathrm{div}_\mu(\mathcal{S}) \equiv \mathrm{div}_\mu(Y_i)Y_i$, so the drift of the $\mathcal{A}$-diffusion decomposes into components associated to antisymmetric and symmetric brackets

$$
\mathrm{d}Z_t = \underbrace{\mathrm{curl}_\mu(\mathcal{A})\,\mathrm{d}t}_{\text{antisymmetric}} + \underbrace{\tfrac{1}{2}\mathrm{div}_\mu(\mathcal{S})\mathrm{d}t}_{\text{symmetric}} + \underbrace{Y_i \circ \mathrm{d}W_t^i}_{\text{noise}}, \tag{4.34}
$$

which for $\mu_H = e^{-\beta H}\mu_{\mathcal{F}}$ becomes

$$
\mathrm{d}Z_t = \Big(X_H^{\mathcal{A}} + \mathrm{curl}_{\mu_{\mathcal{F}}}(\mathcal{A})\Big)\mathrm{d}t + \tfrac{1}{2}\Big(\mathrm{div}_{\mu_{\mathcal{F}}}(\mathcal{S}) - \beta X_H^{\mathcal{S}}\Big)\mathrm{d}t + \sum_{i=1}^N Y_i \circ \mathrm{d}W_t^i.
$$

The following corollary derives a powerful decomposition of the generator which generalises the Euclidean case (see for example [294]), and provides a nice connection with the generalisation of the modular vector field associated to brackets:

Corollary 23. *The generator of the $\mathcal{A}$-diffusion can be written as*

$$
\mathcal{L}f = \underbrace{\phi_\mu^{\mathcal{A}}(f)}_{L^2\text{-antisymmetric}} + \underbrace{\tfrac{1}{2}\phi_\mu^{\mathcal{S}}(f)}_{L^2\text{-symmetric}}. \tag{4.35}
$$

Moreover $\phi_\mu^{\mathcal{A}}$ and $\frac{1}{2}\phi_\mu^{\mathcal{S}}$ are respectively antisymmetric and symmetric in $L^2(\mathcal{F},\mu)$. Hence $\mathcal{L}$ is symmetric iff $\phi_\mu^{\mathcal{A}} = 0$, i.e., μ is a Hamiltonian measure of the $\mathcal{A}$-mechanics. In general, the

generator of (4.24) *satisfies* $\mathcal{L} = \frac{1}{2}\phi_\mu^{\mathcal{S}}$ *iff the* $(\mu, 1)$*-Fokker-Planck current vanishes, in which case we say that* μ *satisfies the* ***detailed balance condition****.*

Proof. The formula for the generator follows from (4.34) and $\mathrm{div}_\mu(X_f^{\mathcal{S}}) = \mathrm{div}_\mu(Y_i)Y_i(f) + Y_iY_if$, as we prove using a local argument: given local coordinates (z^i), writing $\mu = p_\infty \mathrm{d}z$ we have

$$\mathrm{div}_\mu(X_f^{\mathcal{S}}) = \frac{1}{p_\infty}\partial_r(p_\infty S^{jr}\partial_j f) = \frac{1}{p_\infty}\partial_r(p_\infty Y_i^j Y_i^r \partial_j f) = \mathrm{div}_\mu(Y_i)Y_i(f) + \partial_r(Y_i^j\partial_j f)Y_i^r$$

and the result follows from the proof of (4.25). From the proof of (4.26) we know the (formal) adjoint of $\phi_\mu^{\mathcal{A}}$ is $f \mapsto -\mathrm{div}_\mu(f\phi_\mu^{\mathcal{A}})$, and $-\mathrm{div}_\mu(f\phi_\mu^{\mathcal{A}}) = -\phi_\mu^{\mathcal{A}}(f)$ as noted in the proof of proposition 13. From (4.26) we also know that the adjoint of $\phi_\mu^{\mathcal{S}}$ is $f \mapsto -\mathrm{div}_\mu\left(f\mathrm{div}_\mu(Y_i)Y_i\right) + \mathrm{div}_\mu\big(\mathrm{div}_\mu(fY_i)Y_i\big) = \mathrm{div}_\mu\big(Y_i(f)Y_i\big) = \mathrm{div}_\mu\big(\mathcal{S}^\sharp(\mathrm{d}f)\big) = \phi_\mu^{\mathcal{S}}(f)$. ■

The decomposition of the generator above allows us for example to show that for appropriate transformations $S : \mathcal{F} \to \mathcal{F}$ as defined in the next corollary, the generator of the $\mathcal{A}$-diffusion may be shown to be reversible up to S, which extends the reversibility up to momentum flip of the Langevin diffusion in Euclidean space [329, 114].

Corollary 24. *Let* S *be a target-preserving diffeomorphism,* $S_\sharp\mu = \mu$*, which is an* $\mathcal{A}$*-antimorphism, and a* $\mathcal{S}$*-morphism (for example* $S_*Y_i = \pm Y_i$*). Then the generator of the* $\mathcal{A}$*-diffusion* (4.35) *is* ***reversible up to*** S*, that is*

$$\langle f, \mathcal{L}h\rangle_\mu = \langle \mathcal{L}S^*f, S^*h\rangle_\mu .$$

Proof. First note that for any $\mathcal{B}$-Hamiltonian vector field, since $S^*\mu = \mu$, then, using proposition 6.3.5 [3], $S^*\mathcal{L}_{X_f}\mu = \mathcal{L}_{(S^{-1})_*X_f}S^*\mu = \mathcal{L}_{(S^{-1})_*X_f}\mu$, so $\mathrm{div}_\mu(X_f)\circ S = \mathrm{div}_\mu((S^{-1})_*X_f)$. Now consider the antisymmetric part, $\mathcal{L} = \phi_\mu^{\mathcal{A}}$. Then $\langle f, \phi_\mu^{\mathcal{A}}h\rangle_\mu = -\langle \phi_\mu^{\mathcal{A}}f, h\rangle_\mu = -\langle S^*\phi_\mu^{\mathcal{A}}f, S^*h\rangle_\mu$, where we have used $S_\sharp\mu = \mu$ in the last equality. Moreover $S^*\phi_\mu^{\mathcal{A}}f = -\phi_\mu^{\mathcal{A}}S^*f$ since S is an $\mathcal{A}$-antimorphism, $(S^{-1})_*X_f^{\mathcal{A}} = -X_{f\circ S}^{\mathcal{A}}$. The proof of the symmetric part is analogous, except we have $S^*\phi_\mu^{\mathcal{S}}f = \phi_\mu^{\mathcal{S}}S^*f$. Note that if $S_*Y_i = \pm Y_i$ then S is a $\mathcal{S}$-morphism, since $(S^{-1})_*X_f^{\mathcal{S}} = (S^{-1})_*(Y_i(f)Y_i) = Y_i(f)\circ S(S^{-1})_*Y_i = \left(S_*(S^{-1})_*Y_i\right)(f)\circ S(S^{-1})_*Y_i = \left((S^{-1})_*Y_i\right)(S^*f)(S^{-1})_*Y_i = Y_i(S^*f)Y_i = X_{S^*f}^{\mathcal{S}}$. ■

Remark 25. We have thus constructed a natural extension of the Euclidean recipe to arbitrary manifolds, showing it is a direct consequence of the canonical geometry of the target measure. Combining the results from the previous sections, the various terms in the resulting diffusion

equation have the following interpretations:

$$\mathrm{d}Z_t = \overbrace{\underbrace{(\quad \mathrm{curl}_{\mu_{\mathcal{F}}}(\mathcal{A})}_{\text{volume-preserving}} + \underbrace{\qquad X_H^{\mathcal{A}} \qquad}_{\text{density-preserving}})\mathrm{d}t}^{\text{local potential of Fokker–Planck current}} \quad + \quad \overbrace{\mu^\sharp(\gamma)\mathrm{d}t}^{\text{topological obstruction}}}^{\text{conservative } L^2(\mu)\text{-antisymmetric drift}}$$
$$+ \underbrace{(\quad \underbrace{-\tfrac{1}{2}Y_i(H)Y_i}_{\text{density-dissipative drift}} \quad + \overbrace{\underbrace{\tfrac{1}{2}\mathrm{div}_{\mu_{\mathcal{F}}}(Y_i)Y_i}_{\text{volume-dissipative drift}} \quad)\mathrm{d}t + \underbrace{Y_i \circ \mathrm{d}W_t^i}_{\text{Stratonovich noise}}}^{\text{volume-preserving } L^2(\mu_{\mathcal{F}})\text{-symmetric diffusion}}}_{L^2(\mu)\text{-symmetric fluctuation-dissipation balance}}. \tag{4.36}$$

Finally, we observe that double bracket dissipations arise when we choose the noise fields to be $\mathcal{A}$-Hamiltonian vector fields $Y_i \equiv X_{H_i}$ generated by "noise Hamiltonians" $H_1, \ldots, H_N$, for $i = 1, \ldots, N$. In that case the symmetric bracket is related to the antisymmetric bracket $\mathcal{A}$

$$\mathcal{S}(\mathrm{d}f, \mathrm{d}h) = X_{H_i}(f)X_{H_i}(h) = \{H_i, f\}\{H_i, h\}.$$

The $\mathcal{A}$-diffusion then reads

$$\mathrm{d}Z_t = (X_H + \phi_{\mu_{\mathcal{F}}})\,\mathrm{d}t + \tfrac{1}{2}\sum_{i=1}^N (\phi_{\mu_{\mathcal{F}}}(H_i)X_{H_i} - \beta\{H_i, H\}X_{H_i})\,\mathrm{d}t + \sum_{i=1}^N X_{H_i} \circ \mathrm{d}W_t^i,$$

which further simplifies to

$$\mathrm{d}Z_t = X_H\mathrm{d}t - \frac{\beta}{2}\sum_{i=1}^N \{H_i, H\}X_{H_i}\mathrm{d}t + \sum_{i=1}^N X_{H_i} \circ \mathrm{d}W_t^i, \tag{4.37}$$

on a bracket-Hamiltonian manifold ($\phi_{\mu_{\mathcal{F}}} = 0$). The noise component of the generator and Fokker-Planck operators is then in double bracket form

$$\mathcal{L}f = \{H, f\} - \frac{\beta}{2}\sum_{i=1}^N \{H_i, H\}\{H_i, f\} + \tfrac{1}{2}\sum_{i=1}^N \{H_i, \{H_i, f\}\},$$
$$\mathcal{L}^*f = -\{H, f\} + \frac{\beta}{2}\sum_{i=1}^N \{H_i, f\{H_i, H\}\} + \tfrac{1}{2}\sum_{i=1}^N \{H_i, \{H_i, f\}\}.$$

4.5 Langevin Diffusions and Irreversible MCMC algorithms

Consider the second order Langevin diffusion on $\mathbb{R}^n \times \mathbb{R}^n$

$$\mathrm{d}Q_t = \nabla_p H \mathrm{d}t, \qquad \mathrm{d}P_t = -\nabla_q H \mathrm{d}t - \frac{\beta}{2}\gamma(Q_t)\nabla_p H \mathrm{d}t + \lambda(Q_t)\mathrm{d}W_t, \tag{4.38}$$

where γ and λ are matrix fields related through the fluctuation-dissipation relation $\lambda\lambda^\top = \gamma$ [215]. As mentioned previously such Langevin diffusions play a central role in the context of statistical physics and irreversible MCMC, so we now consider the problem of characterising their geometric structure, by specifying their relation to the complete class of measure-preserving diffusions, and generalise the method we proposed in [17] to sample on Lie groups. To do so we consider the following decomposition of the $\mathcal{A}$-diffusion on the tangent bundle $\mathcal{F} \equiv T\mathcal{M}$

$$\mathrm{d}Z_t = \underbrace{\left(X_H + \phi_{\mu_{T\mathcal{M}}}\right)\mathrm{d}t}_{(A)} + \underbrace{\tfrac{1}{2}\sum_{i=1}^{N} \mathrm{div}_{\mu_{T\mathcal{M}}}(Y_i)Y_i \mathrm{d}t}_{(B)} + \underbrace{\sum_{i=1}^{N}\left(Y_i \circ \mathrm{d}W_t^i - \frac{\beta}{2}Y_i(H)Y_i\right)}_{(C)}. \tag{4.39}$$

The characterising property of Langevin processes is that they are measure-preserving, thus must be $\mathcal{A}$-diffusions, but also contain no rotationnel term associated to the reference measure. Moreover underdamped ones, as in (4.38), have the additional property that the stochastic component only acts on momentum $\mathrm{d}P_t$: the noise is vertically directed. Thus, to locally recover Langevin diffusions we will need to choose our $\mathcal{A}$-mechanics, and reference measure $\mu_{T\mathcal{M}}$ s.t., **(i)** $\phi_{T\mathcal{M}} = 0$, and **(ii)** (B) vanishes whenever the noise are vertical vector fields (in §4.6.3 we will characterise P-preserving diffusions for which (i) holds and (B) vanishes when $\mathcal{F}$ is arbitrary). Such requirements will be satisfied for pairs $(\mathcal{A}, \mu_{T\mathcal{M}})$ that we will refer to as **Langevin pairs**. As we will then see, if $(\mathcal{A}, \mu_{T\mathcal{M}})$ is a Langevin pair, the term (C) turns into a OU process whenever H is a simple mechanical system $\pi^*V + \frac{1}{2}\|\cdot\|^2$, and the noise fields (Y_i) points only along the vertical direction. In particular we will prove that the Riemann–Newton pair $(\omega_\flat, \omega_\flat^n)$ is Langevin, and targets of the form $e^{-\pi^*V - \frac{1}{2}\|\cdot\|^2}\omega_\flat^n$ naturally arise when the target P is expressed in terms of the Riemannian one-density on $\mathcal{M}$. Hence, for such systems (4.39) nicely decomposes into measure-preserving dynamics + OU steps, which gives rise to the class of Langevin-based irreversible MCMC. In the presence of symmetries we may further pullback $\omega_\flat$ and work in the reduced symplectic space.

In the following we denote by $\mathcal{V} \in \mathfrak{X}(T\mathcal{M})$ the canonical vector field defined by the vertical lift, $\mathcal{V} : v_q \mapsto \mathrm{ver}_{v_q} v_q$.

4.5.1 Langevin Pairs and Riemannian Newton Diffusions

We have developed a rigorous theory of term (A) during our analysis of HMC desideratum §4.3.1, where we showed it was $e^{-H}\mu_{\mathcal{F}}$-preserving. Moreover we have seen that the vanishing of the curl vector field $\phi_{\mu_{T\mathcal{M}}}$ was equivalent to the reference measure being Hamiltonian for the $\mathcal{A}$-mechanics, that is $(\mathcal{A}, \mu_{T\mathcal{M}})$ form an $\mathcal{A}$-Hamiltonian system (in which case X_H is $e^{-H}\mu_{\mathcal{F}}$-preserving). To construct vertical noise fields we will choose vector fields $X_i \in \mathfrak{X}(\mathcal{M})$: then the lifts $Y_i \equiv \mathcal{V} \circ X_i \circ \pi \in \mathfrak{X}(T\mathcal{M})$ are vertical vector fields which only points along the vertical direction as required. On the other hand the term (B) is still typically non-zero and intractable (it also does not preserve any measure of interest[17]). This leads to the definition of a **Langevin pair**, which is a pair $(\mathcal{A}, \mu_{T\mathcal{M}})$ s.t.,

- $(\mathcal{A}, \mu_{T\mathcal{M}})$ is a $\mathcal{A}$-Hamiltonian system,
- $\mu_{T\mathcal{M}}$ is **horizontal**, i.e., $\mathcal{L}_Y(\mu_{T\mathcal{M}}) = 0$ for any vertical vector field Y.

The $\mathcal{A}$-diffusion (4.28) generated by a Langevin pair $(\mathcal{A}, \mu_{T\mathcal{M}})$ and vertical noise fields (Y_i) locally yields the Langevin diffusions (4.38), with the deterministic Hamiltonian mechanics therein replaced by a more general $\mathcal{A}$-mechanics. Moreover, by theorem 17, terms (A) and (C) are now independently measure-preserving, which motivates the use of splitting methods for the numerical implementation.

As we will see below, choosing the noise to be vertical ensures that when $(\mathcal{A}, \mu_{T\mathcal{M}})$ is a Langevin pair and the Hamiltonian corresponds to a simple mechanical system, $H \equiv \pi^* V + \frac{1}{2}\|\cdot\|^2$, then (C) in (4.39) turns into a OU process on the fibre, for which there is a tractable solution, and which preserves the Gaussian distribution on the fibre defined by the Riemannian metric. An important example of Langevin pair and target $\mu_H \equiv e^{-\pi^* V - \frac{1}{2}\|\cdot\|^2}\mu_{T\mathcal{M}}$ arises when the target P on $\mathcal{M}$ is expressed in terms of the Riemannian measure, $P = e^{-V}\mathrm{vol}$. In that case using the Gaussian distributions $\mathcal{N}(0, \langle\cdot,\cdot\rangle^{-1})$ as a probability disintegration ζ, the lift of P to the Riemannian Newton space $(T\mathcal{M}, \omega_\flat, \omega_\flat^n)$ with respect to the disintegration yields the canonical distribution $\mu_H \equiv \pi^* P \wedge \zeta = e^{-H}\omega_\flat^n$ where the Hamiltonian takes the form $H \equiv \pi^* V + \frac{1}{2}\|\cdot\|^2$ (see §3.3.2). As the following theorem shows, $\omega_\flat^n$ is indeed horizontal, and thus the $\omega_\flat^{-1}$-diffusion generated by μ_H decomposes into a μ_H-preserving Hamiltonian flow (A), and a μ_H-preserving OU process (C) on the fibre whenever the noise fields (Y_i) are chosen to be vertical:

Theorem 26. *Suppose $(\mathcal{A}, \mu_{T\mathcal{M}})$ is a Langevin pair. If we choose the noise fields to be the vertical fields $Y_i \equiv \mathcal{V} \circ X_i \circ \pi$ for $X_i \in \Gamma(T\mathcal{M})$, then the $\mathcal{A}$-diffusion generated by $e^{-H}\mu_{T\mathcal{M}}$,*

[17] See [53] for a discussion of the simulation of (4.39) in Euclidean space when the terms (B-C) are present but (A) vanishes.

with $H \equiv \pi^* V + \frac{1}{2}\|\cdot\|^2$, *is*

$$\mathrm{d}(q_t, v_t) = \underbrace{X_H(q_t, v_t)\mathrm{d}t}_{\mathcal{A}-\textit{Mechanics}} - \underbrace{\frac{\beta}{2}\left\langle X_i(q_t), v_t\right\rangle_{q_t} \mathcal{V}(X_i(q_t))\mathrm{d}t}_{\textit{vertical kinetic Dissipation}} + \underbrace{\mathcal{V}(X_i(q_t)) \circ \mathrm{d}W_t}_{\textit{vertical randomness}}, \tag{4.40}$$

or in tangent-lifted coordinates

$$\mathrm{d}(q_t, v_t) = \Big(X_H(q_t, v_t) - \frac{\beta}{2} M(q_t) v_t\Big)\mathrm{d}t + \sigma(q_t) \circ \mathrm{d}W_t, \tag{4.41}$$

where $M(q)v = M(q)_{rj} v^j \partial_{v^r} \equiv (\sigma\sigma^\top G(q))_{rj} v^j \partial_{v^r}$ *and* $\sigma \equiv \sigma_{ji}\partial_{v^j} \equiv (X_i)^j \partial_{v^j}$, *and* $G_{ij}(q) = \langle \partial_{x^i}, \partial_{x^j}\rangle_q$. *In particular* $(\omega_\flat^{-1}, \omega_\flat^n)$ *is a Langevin pair. Moreover the Hamiltonian noise fields* $Y_i \equiv X_{U_i \circ \pi}$ *for* $U_i : \mathcal{M} \to \mathbb{R}$, *are vertical,* $Y_i = -\mathcal{V} \circ \nabla U_i \circ \pi$, *and* $Y_i(H) = -\mathrm{d}U_i \circ \partial\pi(X_T)$.

Proof. If $X \in \Gamma(T\mathcal{M})$ and $\mathcal{V} \in \Gamma(T\mathcal{M}, TT\mathcal{M})$ is the canonical vector field, $\mathcal{V} : v_q \mapsto \mathrm{ver}_{v_q} v_q$,[18] locally $\mathcal{V} = v^k \partial_{v^k}$, $X = X^k(x)\partial_{x^k}$, so their composition $\mathcal{V} \circ X \circ \pi \in \Gamma(T\mathcal{M}, TT\mathcal{M})$ is $\mathcal{V} \circ X \circ \pi(x, v) = X^k(x)\partial_{v^k}$. Then

$$\begin{aligned} Y_i(H)(x, v) &= X_i^k(x)\partial_{v^k}\big(V(x) + T(x, v)\big) = X_i^k(x)\partial_{v^k}\big(\tfrac{1}{2} G_{rj}(x) v^r v^j\big) \\ &= X_i^k(x) G_{jk}(x) v^j = \left\langle X_i(x), v\right\rangle_x. \end{aligned}$$

Set $\sigma_{ji} \equiv X_i^j$, so $Y_i(x, v) = \sigma_{ji}(x)\partial_{v^j}$. We have

$$\begin{aligned} Y_i(H) Y_i(x, v) &= \left\langle X_i(x), v\right\rangle_x \mathrm{ver}_{X_i(x)}(X_i(x)) = X_i^k(x) G_{jk}(x) v^j X_i^r(x)\partial_{v^r} \\ &= \sigma_{ki}(x) G_{jk}(x) v^j \sigma_{ri}(x)\partial_{v^r} = \big(\sigma\sigma^\top(x)\big)_{rk} G_{kj}(x) v^j \partial_{v^r}, \end{aligned}$$

and (4.41) and its local expression follow. Moreover the symplectic measure is indeed horizontal (this fact may be traced back to the fact that the Liouville 1-form is horizontal). Indeed, in tangent-lifted coordinates $\omega_\flat^n = |G|\mathrm{d}x \wedge \mathrm{d}v$, so locally the divergence of $Z \in \Gamma(T\mathcal{M}, TT\mathcal{M})$ is $\mathrm{div}_{\omega_\flat}(Z) = \frac{1}{|G|}\partial_{x^i}\big(|G| Z^i\big) + \partial_{v^j}\overline{Z}^j$, and from the previous local expressions we see $\mathrm{div}_{\omega_\flat}(Y_i) = 0$, and thus $\omega_\flat^n$ is horizontal. Finally if $U_i : \mathcal{M} \to \mathbb{R}$ and $Y_i \equiv X_{U_i \circ \pi} = -\mathcal{V} \circ \nabla U_i \circ \pi$, then $Y_i(H) = \omega_\flat\big(X_H, X_{U_i \circ \pi}\big) = -X_H(U_i \circ \pi) = -\mathrm{d}U_i \circ \partial_\pi(X_H) = -\mathrm{d}U_i \circ \partial_\pi(X_T)$ since X_V is vertical. ∎

When $E \to \mathcal{M}$ is isomorphic to $T\mathcal{M}$ in the category of vector bundles over $\mathcal{M}$, for example $E = T^*\mathcal{M}$, then we can use the isomorphism to rewrite (4.40) over E. In particular when $\mathcal{M} = \mathbb{R}^n$ and $T(q, v) = \frac{1}{2} v^\top G v$, using the musical isomorphism the SDE (4.41) becomes the

[18] Note $v_q \mapsto \mathrm{ver}_{v_q} v_q \in TT\mathcal{M}$, $\mathrm{ver}_{v_q} v_q : f \mapsto \frac{\mathrm{d}}{\mathrm{d}t} f(q, v_q + t v_q)|_{t=0}$ for any $f \in C^1(T\mathcal{M})$. Locally this is just the directional derivative of f at (q, v_q) in the direction $(0, v_q)$, $(0, v) \cdot \nabla f(q, v) = v^k \partial_{v^k} f(q, v)$, or $\mathcal{V} = v^k \partial_{v^k}$.

Langevin dynamics on phase space used in statistical physics (with $\lambda \equiv G\sigma$, $\gamma \equiv G\sigma\sigma^\top G$)

$$\mathrm{d}Q_t = G^{-1}P_t\mathrm{d}t, \qquad \mathrm{d}P_t = -\nabla V(Q_t)\mathrm{d}t - \frac{\beta}{2}\gamma(Q_t)G^{-1}(Q_t)P_t\mathrm{d}t + \lambda(Q_t)\mathrm{d}W_t,$$

where now the definition $M \equiv \sigma\sigma^\top G$ plays the role of the fluctuation-dissipation relation which ensures the target $\propto e^{-H}\mathrm{d}q\mathrm{d}p$ is preserved. We thus see that, as claimed, (4.40) is the manifold generalisation of the usual Langevin SDE, where the noise vector fields X_i represent the columns of the "vertical matrix" σ, which only introduces randomness along the velocity, and $\langle X_i(q_t), v_t\rangle_{q_t}$ describes the rate of change of the kinetic energy along the noise. On $\mathcal{F} = \mathbb{R}^n \times \mathbb{R}^n$ another example of Langevin pair consists of choosing $\mathcal{A}$ to be a constant antisymmetric matrix (which is Poisson but not necessarily symplectic), and $\mu_{\mathcal{F}}$ the Lebesgue measure $\mathrm{d}q\mathrm{d}p$. It would be interesting to analyse the optimal properties of a subclass of these of Langevin processes as proposed in [213].

4.5.2 iLMCMC Algorithm

We can now proceed in a similar fashion to [289] to build irreversible MCMC schemes. Indeed, by using vertical noise fields we see that (4.40) naturally decomposes into a Hamiltonian part and a vertical part which remains within the initial fibre, and thus only shifts the velocity (i.e., replaces the HMC heat bath). We split (4.40) into a $\mathcal{A}$-Hamiltonian part on $T\mathcal{M}$

$$\frac{\mathrm{d}z}{\mathrm{d}t} = X_H(z), \tag{4.42}$$

and a OU process within the tangent fibres $T_q\mathcal{M}$ (since the vertical lift is an isomorphism $T_v(T_q\mathcal{M}) \cong T_q\mathcal{M}$)

$$\dot{q} = 0, \qquad \mathrm{d}v_t = -\frac{\beta}{2}\langle X_i(q), v_t\rangle_q X_i(q)\mathrm{d}t + X_i(q) \circ \mathrm{d}W_t,$$

which locally has the form[19]

$$\dot{q} = 0, \qquad \mathrm{d}v_t = -\frac{\beta}{2}M(q)v_t\mathrm{d}t + \sigma(q) \circ \mathrm{d}W_t, \tag{4.43}$$

and preserves the Gaussian $\propto e^{-\frac{1}{2}\beta\langle v,v\rangle_q}\mathrm{d}v = \mathcal{N}(0, G^{-1}(q)/\beta)$. Thus by choosing vertical noise fields, and a Langevin pair in the general diffusion (4.28) we obtain a diffusion which naturally split into a $\mathcal{A}$-Hamiltonian mechanics, and a tractable OU-process as in the Euclidean case. Below we briefly consider the $\mathcal{A} = \omega_\flat^{-1}$-case and its geodesic integration, and then then case of

[19] with $T_q\mathcal{M} \cong \mathbb{R}^n$ using the local basis, $M(q) \equiv \sigma(q)\sigma^\top(q)G(q)$ where $\sigma_{ji}(q) \equiv (X_i(q))^j$.

Lie groups.

4.5.2.1 iLMCMC with Geodesic Integrators on Embedded Manifolds

Suppose now that $\iota : \mathcal{M} \hookrightarrow \mathbb{R}^k$ is an embedded manifold with Riemannian metric defined by restricting the Euclidean dot product. We assume we have a tractable expression for the geodesic flow Φ^T, a C^1-extension W of the potential energy V in the coordinates of the embedding, and that the Riemannian metric corresponds to the one we used to define the reference measure of P - else we need to add a Radon-Nykodym term in our Hamiltonian, or use the Riemannian integrator [210]. Then we use a geodesic integrator which approximates (4.42) by a composition of geodesic flow Φ^T, and vertical gradient flow Φ^V whose integral curve starting from $(q_0, v_0) \in T\mathbb{R}^k|_{\mathcal{M}}$ is

$$q(t) = q_0 \qquad v(t) = v_0 - t\mathrm{hor}\,(\nabla_{q_0} W)$$

where $\nabla_{q_0} W$ is the Euclidean gradient and hor the horizontal projection onto the tangent space of $\mathcal{M}$ (recall §3.6). Similarly to implement (4.43) we need the noise vector fields X_i to be expressed in the coordinates of the embedding: that is each noise field X_i is given by a vector field $b_i = (b_i^1, \ldots, b_i^k)$ on $T\mathbb{R}^k|_{\mathcal{M}}$ (i.e., $\partial\iota \circ X_i = b_i \circ \iota$). Then b_i defines the i^{th} column of the matrix σ and the process (4.43) is $\partial_q\iota$-related to the following process on $\mathbb{R}^k$ ("$\cdot$" is the dot product)

$$\mathrm{d}y_t = -\frac{\beta}{2} b_i(q) \cdot y_t b_i(q)\mathrm{d}t + \sigma(q)\mathrm{d}W_t.$$

Note that our potential energy does not include a $\log\det G$ term, unlike the geodesic MCMC of [225]. Indeed, as we discussed in the previous chapter, this terms does not give rise to a meaningful potential energy on a manifold and should not be present. In the very special case in which $\mathcal{M} = \mathbb{R}^k$ and we have fixed a coordinate system (so $\mathcal{M}$ is no longer a manifold), then we can add the "correction term" $\log\det G$ to the potential energy in order to ensure the algorithm generates samples from $e^{-V}\mathrm{d}q$ rather than $e^{-V}\mathrm{vol}$, since typically distributions on Euclidean space are expressed in terms of the Lebesgue measure (see also [160], §5.1.4).

4.5.2.2 iLMCMC on Lie Groups

Suppose now that the configuration space $\mathcal{M} = \mathcal{G}$ is a Lie group with left-invariant metric (the identity of $\mathcal{G}$ is denoted 1), so that the Riemannian Newton space can be identified with the symplectic manifold $\mathcal{G} \times \mathfrak{g}$ by left-trivialisation (see §3.9.1). Let $v^i : \mathfrak{g} \to \mathbb{R}$, $v^i(\xi) = \theta_1^i(\xi)$ be the coordinates on $\mathfrak{g}$ associated to a basis ξ_i. Since $T(\mathcal{G} \times \mathfrak{g}) = T\mathcal{G} \oplus T\mathfrak{g}$, the vector fields on

$\mathcal{G} \times \mathfrak{g}$ can be expanded as $X = a^i e_i + b^i \partial_{v^i}$, where $\partial_{v^i} \in \Gamma(T\mathfrak{g})$. In [17] we chose Hamiltonian fields of the form $Y_i \equiv X_{U_i \circ \pi}$ for some noise potentials $U_i : \mathcal{G} \to \mathbb{R}$. Instead here we will choose the noise fields to be $Y_i = \partial_{v^i}$ in order for the OU process to be computationally cheaper (we could also have $Y_i = f_i(g)\partial_{v^i}$), and from theorem 26 $\operatorname{div}_{\omega^n} Y_i = 0$ since $\partial_{v^i} = \mathcal{V}(\xi_i)$.[20] Then using a kinetic energy T on $\mathfrak{g}$ associated to an inner product (i.e., a left-invariant metric) with matrix G, $T \equiv \frac{1}{2} G_{ij} \theta_1^i \otimes \theta_1^j$, (4.41) becomes

$$\underbrace{\dot{q} = v_t^j e_j(q_t)}_{\text{reconstruction}}, \quad \mathrm{d}v_t = \underbrace{\operatorname{ad}^\top_{v_t} v_t \mathrm{d}t}_{\text{Euler-Arnold}} \underbrace{-G^{jk} e_j(V)(q_t)\xi_k \mathrm{d}t}_{\text{Potential flow}} \underbrace{-\frac{\beta}{2} v_t \mathrm{d}t + \mathrm{d}W_t}_{\text{OU}}, \tag{4.44}$$

where $(q_t, v_t) \in \mathcal{G} \times \mathfrak{g}$. When the inner product on $\mathfrak{g}$ is ad-invariant the Euler-Arnold term vanishes. On SU(3) the diffusion splits into the transition steps used in Hybrid Monte Carlo simulation for lattice QCD, with the OU process replacing the momentum heat bath. In Euclidean space, $e_j = \xi_j = \partial_{x^j}$, and we recover the second order Langevin equation [289, 100].

The Ornstein-Uhlenbeck process $\mathrm{d}v_t = -\frac{\beta}{2} v_t \mathrm{d}t + \mathrm{d}W_t$ on $\mathfrak{g}$ has an explicit solution given by

$$v_{t+h} = e^{-\frac{\beta}{2}h} v_t + \int_t^{t+h} e^{-\frac{\beta}{2}(t+h-s)} \mathrm{d}W_s, \tag{4.45}$$

with transition probability

$$p(v_0, v) = \sqrt{\frac{\beta}{(2\pi)^n (1 - e^{-\beta h})}} \exp\left(-\frac{\beta}{2(1 - e^{-\beta h})} \left\| v - e^{-\frac{\beta}{2}h} v_0 \right\|^2 \right). \tag{4.46}$$

The corresponding irreversible MCMC algorithm is given as follows:

1. Solve OU process exactly until time h by sampling

$$v^* \simeq \mathcal{N}\left(e^{-\frac{\beta}{2}h} v_0, \frac{1}{\beta}(1 - e^{-\beta h})\mathrm{Id} \right), \tag{4.47}$$

   to obtain $(\bar{g}_0, \bar{v}_0) = (g_0, v^*)$;

[20] In fact we can check this directly without relying on local coordinates: writing the left Haar measure as $\Theta \equiv \theta^1 \wedge \cdots \wedge \theta^n$, we have

$$\mathcal{L}_{\partial_{v^i}} \omega^n = \left(\mathcal{L}_{\partial_{v^i}} \mathrm{d}v\right) \wedge \pi^* \Theta + \mathrm{d}v \wedge \left(\mathcal{L}_{\partial_{v^i}} \pi^* \Theta\right).$$

Now $\mathcal{L}_{\partial_{v^i}} \mathrm{d}v^j = \mathrm{d}\mathcal{L}_{\partial_{v^i}} v^j = \mathrm{d}i_{\partial_{v^i}} \mathrm{d}v^j = \mathrm{d}\delta_{ji} = 0$. Then we have $\mathcal{L}_{\partial_{v^i}} \pi^* \Theta = 0$, since the flow Φ of ∂_{v^i} is only non-trivial in the vertical direction. Indeed its flow is $\Phi_t(g, v) = (g, v^1, \ldots, v^i + t, \ldots, v^n)$ so

$$\left(\mathcal{L}_{\partial_{v^i}} \pi^* \Theta\right)(g, v) = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} (\Phi_t^*(g, v) \pi^* \Theta) = \frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=0} \Big(\left(\pi \circ \Phi_t(g, v)\right)^* \Theta \Big) = 0,$$

since $\pi \circ \Phi_t(g, v) = g$ is independent of t.

2. Solve the first-order Euler–Arnold equation, and approximate the Hamiltonian system using N leapfrog trajectories with step size $\delta t > 0$. For example, if the Riemannian metric is bi-invariant, starting at $(\bar{g}_0, \bar{v}_0) = (g_0, v^*)$:

   For $k = 0, \ldots, N-1$: [21]

$$\begin{aligned}
\bar{v}_{k+\frac{1}{2}} &= \bar{v}_k - \frac{\delta t}{2} \mathrm{Tr}\left(\partial_x V^T \bar{g}_k \xi_i\right) \xi_i \\
\bar{g}_{k+1} &= \bar{g}_k \exp\left(\delta t\, \bar{v}_{k+\frac{1}{2}}\right) \\
\bar{v}_{k+1} &= \bar{v}_{k+\frac{1}{2}} - \frac{\delta t}{2} \mathrm{Tr}\left(\partial_x V^T \bar{g}_{k+1} \xi_i\right) \xi_i
\end{aligned}$$

   to obtain $(\bar{g}_N, \bar{v}_N)$.

3. Accept or reject the proposal by a Metropolis-Hastings step. We accept the proposal $(\bar{g}_N, \bar{v}_N)$ with probability

$$\alpha = \min\left\{1, \exp\left(-H(\bar{g}_N, \bar{v}_N) + H(\bar{g}_0, \bar{v}_0)\right)\right\},$$

   and set $(g_1, v_1) = (\bar{g}_N, \bar{v}_N)$. On the other hand, if the proposal is rejected, we set $(g_1, v_1) = (\bar{g}_0, -\bar{v}_0)$.

Compared to the algorithm we had presented in [17], our choice of noise field $Y_i \equiv \partial_{v_i}$ as opposed to $Y_i \equiv X_{U_i \circ \pi}$ avoids having to compute matrix exponentials in the first step of the algorithm (which appears in the latter situation in the form $e^{-\frac{\beta}{2} Dh}$, where D is a matrix given by $D_i^j = Y_i^k Y_k^j$), thus significantly reducing the computational cost. On the other hand choosing noise fields of the form $Y_i = X_{U_i \circ \pi}$ may be useful when the potential energies U_i are adapted to the target distribution, by for example increasing the contribution of the noise at suited locations.

4.6 The Potential Theory of Measures and Complete Recipe of Measure-Constraints Preserving Flows

In this chapter we have shown that any smooth positive target measure P induces a boundary operator and homology groups that are canonically isomorphic to the de Rham cohomology ones. We have shown that this isomorphism allows us to easily transfer old results from differ-

[21] For a non-matrix group, simply replace $\mathrm{Tr}\left(\partial_x V^T \bar{g}_k \xi_i\right)$ with $e_i|_g(V)$.

ential geometry to the statistical context, and in particular proved that the complete recipes of P-preserving diffusions (see §4.4.2) and P-preserving $\mathcal{A}$-mechanics (see §4.3.1.3) were a direct consequence of this canonical isomorphism. Crucially, none of these characterisation rely on additional structures, such as a Riemannian or symplectic structures, or the use of local coordinates. Indeed, the use of such structures is probably the main reason why, despite its importance, the curl_P operator is almost never mentioned in the literature. This is a consequence of the fact that Riemannian and symplectic brackets, which are central in mathematical physics, are non-degenerate brackets (recall §4.2.2.3) that not only induce a measure (the Riemannian and symplectic measures), allowing properties of the induced measures to be derived via properties of the brackets, but also additional structures on the manifold, such as an isomorphism $\mathcal{B}^\sharp$ between $T^*\mathcal{M}$ and $T\mathcal{M}$, special classes of coordinates, and even a compatible connection that allows us to differentiate many tensor fields in the case of Riemannian metrics. These additional structures hide the true role played by the measure itself, which only allows us to identify twisted ℓ-forms with $(n-\ell)$-vector fields, and differentiate multi-vector fields. Yet, as we have shown, this weaker structure is often enough for statistical purposes, wherein the target measure P is the only object we are provided with.

Aside from mathematical clarity and the ability to canonically derive the above recipes on arbitrary manifolds, understanding the intrinsic geometry of P allows us to derive a genuine potential theory of P-preserving flows, wherein multi-vector fields play the analogous role that the potential energy and magnetic vector potential play in physics. We discuss this in §4.6.1. Using the potential theory we can easily construct flows that preserve both a measure and a set of constraints that can be recognised as particular instances of Nambu mechanics. In §4.6.2 we prove that Nambu mechanics gives us an elegant interpretation of the identification between geometry and statistics provided by the P-musical isomorphism. Moreover, using the foliated de Rham cohomology we derive the complete recipe of measure-preserving flows that preserve a given set of constraints. In §4.6.3 we discuss the history of measure-preserving diffusions, and in §4.6.4 we explain why the combination of theoretical and tractability guarantees make Hamiltonian mechanics a particularly powerful flow in the context of sampling.

While in this thesis we have been mainly interested in applying geometrical methods to study measure-preserving flows, in §6.9 we will show that the canonical geometry of P we have derived has implications that go way beyond the context of flows, by explaining the key role it plays in understanding Stein operators. Indeed, these are nothing else than the divergence operator restricted to the image of the P-musical isomorphism applied to the space of $(n-1)$-twisted forms that satisfy Stokes' theorem.

4.6.1 Potential Theory of Measures

In classical mechanics and electromagnetism, any conservative force F and electromagnetic 2-form $\mathbf{F}$ can be expressed in terms of a potential energy V and a vector potential $\mathbf{A}$ respectively

$$F = -\mathrm{d}V, \qquad \mathbf{F} = \mathrm{d}\mathbf{A}.$$

Conversely, it is often convenient to start by constructing potentials with suitable properties in order for these properties to automatically transfer to the induced field. Globally curled multi-vector fields play the analogous role in statistics, and in the previous sections we have seen that P-preserving flows X and P-preserving brackets $\mathcal{B}$ were, at least locally, of the form

$$X = \mathrm{curl}_P(\mathcal{A}), \qquad \mathcal{B} = \mathrm{curl}_P(\mathcal{W}),$$

for some *potentials* $\mathcal{A} \in \mathfrak{X}^2(\mathcal{M})$ and $\mathcal{W} \in \mathfrak{X}^3(\mathcal{M})$. As in physics, we may then be interested in exploiting suitable potentials in order to construct P-preserving multi-vector fields that have desirable properties. For example, the following theorem shows the P-rotationnel of an invariant measure preserves the invariance, which provides us with a simple recipe to construct measure-preserving fields that respect the symmetry of the measure, by means of invariant higher-order multi-vector fields.

Theorem 27. *Suppose P is invariant, i.e., $\mathcal{L}_X P = 0$ for some vector field X. If $\mathcal{V} \in \mathfrak{X}^k(\mathcal{M})$ is invariant, then $P^\flat(\mathcal{V})$, and if $\alpha \in \Omega^k(\mathcal{M})$ is invariant, then $P^\sharp(\alpha)$ is invariant. Hence $curl_P(\mathcal{V})$ is invariant.*

Proof. First, note that $P^\flat(\mathcal{V})$ is invariant since

$$\begin{aligned}
\mathcal{L}_X P^\flat(\mathcal{V}) = \mathcal{L}_X i_\mathcal{V} P &= (\mathrm{d} \circ i_X + i_X \circ \mathrm{d}) \circ i_\mathcal{V} P = (\mathrm{d} i_{\mathcal{V} \wedge X} + i_X \circ \mathrm{d} \circ i_\mathcal{V})\, P \\
&= \left(i_\mathcal{V} \circ \mathrm{d} \circ i_X + (-1)^k i_{\mathcal{V} \wedge X} \circ \mathrm{d} - i_{[\mathcal{V},X]}\right) P \\
&= \left(i_\mathcal{V} \circ \mathrm{d} \circ i_X + i_\mathcal{V} \circ i_X \circ \mathrm{d} - i_{[\mathcal{V},X]}\right) P \\
&= \left(i_\mathcal{V} \circ \mathcal{L}_X - i_{[\mathcal{V},X]}\right) P = (-1)^{k+1} P^\flat(\mathcal{L}_X \mathcal{V}) = 0.
\end{aligned}$$

Now if $\mathcal{L}_X \alpha = 0$ then, using the result above

$$0 = \mathcal{L}_X P^\flat \circ P^\sharp \circ \alpha = \mathcal{L}_X P^\flat \left(P^\sharp(\alpha)\right) = \mathcal{L}_X i_{P^\sharp(\alpha)} P = (-1)^{n-k+1} P^\flat \left(\mathcal{L}_X P^\sharp(\alpha)\right),$$

so $P^\flat \left(\mathcal{L}_X P^\sharp(\alpha)\right) = 0$ which implies $\mathcal{L}_X P^\sharp(\alpha) = 0$, i.e., $P^\sharp(\alpha)$ is invariant. Finally, if $\mathcal{V}$ is invariant, then $P^\flat(\mathcal{V})$ is invariant, thus $\mathrm{d}P^\flat(\mathcal{V})$ is invariant (since $\mathcal{L}_X \circ \mathrm{d}P^\flat(\mathcal{V}) = \mathrm{d} \circ i_X \circ \mathrm{d}P^\flat(\mathcal{V}) = \mathrm{d} \circ \mathcal{L}_X P^\flat(\mathcal{V}) = 0$), thus $\mathrm{curl}_P(\mathcal{V}) = P^\sharp \circ \mathrm{d} \circ P^\flat(\mathcal{V})$ is invariant. ∎

Example 28. *That $curl_P$ maps invariant tensors to invariant tensors is in fact a special case of the following characterisation, which follows from $[X, B] = \mathcal{L}_X B$ and the fact the rotationnel is (up to a sign) a derivation of the Schouten bracket [102, Sec. 2.5]:*

$$\mathcal{L}_X \, curl_P B = curl_P(\mathcal{L}_X B) - i_{\mathrm{d}\, div_P(X)} B.$$

In particular if $div_P X$ is locally constant then

$$\mathcal{L}_X \, curl_P B = curl_P(\mathcal{L}_X B),$$

and $\mathcal{L}_X \, curl_P B = 0$ is equivalent to $\mathcal{L}_X B$ $curl_P$-free. As an example, take $\mathcal{V} = x\partial_x \wedge \partial_y \wedge \partial_z$. Then $\mathcal{L}_{\partial_x} \mathcal{V} = \partial_x \wedge \partial_y \wedge \partial_z \neq 0$, yet $curl_{\mathbf{dx}} \mathcal{V}$ is invariant, and indeed we have $curl_{\mathbf{dx}}(\mathcal{L}_{\partial_x} \mathcal{V}) = 0$. The canonical Poisson bi-vector field $\Pi = \partial_{q^i} \wedge \partial_{p_i}$, which is constant, can be constructed as the $curl_{\mathbf{dqdp}}$ of a linear combination of 3-vector fields of the form $z^j \partial_{z^j} \wedge \partial_{q^i} \wedge \partial_{p_i}$.

Occasionally the P-preserving transformations we are interested in are not associated to the flow of a vector field, an example of which was given in §4.3.3 wherein we were concerned with P-preserving bracket anti-morphisms S, in order to construct S-reversible splitting integrators. Specifically, given a P-preserving diffeomorphism S, we were interested in constructing a bracket $\mathcal{A}$ s.t., $\mathrm{curl}_P(\mathcal{A}) = 0$ and $S_* \mathcal{A} = -\mathcal{A}$. The following theorem shows that this may be achieved by using 3-vector fields $\mathcal{W}$ s.t., $S_* \mathcal{W} = -\mathcal{W}$, and applying the rotationnel.

Theorem 29. *Let $F : \mathcal{N} \to \mathcal{M}$ be a diffeomorphism, and denote by $F_* : \mathfrak{X}^\ell(\mathcal{N}) \to \mathfrak{X}^\ell(\mathcal{M})$ the pushforward of multi-vector fields. Then*

$$(F^* P)^\flat = F^* \circ P^\flat \circ F_*, \qquad (F^* P)^\sharp = (F^{-1})_* \circ P^\sharp \circ (F^{-1})^*.$$

Hence

$$curl_{F^* P} = (F^{-1})_* \circ curl_P \circ F_*.$$

Proof. Since the pullback of tensor field commutes with contractions and tensor products we have (denoting the right interior product contraction by $\mathcal{C}$), $(F^*P)^\flat(\mathcal{V}) = i_{\mathcal{V}} F^* P = \mathcal{C}(F^* P \otimes \mathcal{V}) = \mathcal{C}(F^* P \otimes F_*^{-1} F_* \mathcal{V}) = \mathcal{C}(F^* P \otimes F^* F_* \mathcal{V}) = \mathcal{C} \circ F^*(P \otimes F_* \mathcal{V}) = F^* \mathcal{C}(P \otimes F_* \mathcal{V}) = F^* P^\flat(F_* \mathcal{V})$. Thus $(F^*P)^\sharp = (F^{-1})_* \circ P^\sharp \circ (F^{-1})^*$ since, $(F^*P)^\sharp \circ (F^*P)^\flat = \mathrm{Id}$, and the final result follows because the pullback commutes with the exterior derivative. ∎

Thus, if $S_* P = P$ and $S_* \mathcal{W} = -\mathcal{W}$, then $S_* \mathrm{curl}_P(\mathcal{W}) = \mathrm{curl}_P(S_* \mathcal{W}) = -\mathrm{curl}_P(\mathcal{W})$, so that $\mathcal{A} \equiv \mathrm{curl}_P(\mathcal{W})$ has the desired properties.

Another class of flows which are fundamental in statistics and physics (for example in the context of molecular systems or sampling on submanifolds) are those that preserve both a mea-

sure and a set of constraints. The simplest example of such flows is of course the Hamiltonian vector field, which preserves both the energy H and the symplectic measure. In (4.17), using the commuting properties of the interior product with exact forms and the rotationnel, we have seen that such flows were constructed by applying the rotationnel to bi-vector fields of the form $i_{\mathrm{d}H}\mathcal{V}$. More generally, we have the following

Corollary 30. *Let $\gamma \equiv \mathrm{d}H_1 \wedge \cdots \wedge \mathrm{d}H_\ell$ be the ℓ-form associated to (usually independent) constraints $\{H_i\}$. For any $\mathcal{Z} \in \mathfrak{X}^{3+\ell}(\mathcal{M})$ and $\mathcal{V} \in \mathfrak{X}^{2+\ell}(\mathcal{M})$, let*

$$\mathcal{A} \equiv curl_P(i_\gamma \mathcal{Z}) \qquad and \qquad X \equiv curl_P(i_\gamma \mathcal{V}). \tag{4.48}$$

Then the $\mathcal{A}$-mechanics $\mathfrak{X}_{\mathcal{A}}(\mathcal{M})$ and the dynamics X preserve both P and $\{H_i\}$, that is for any H_i

$$\mathcal{L}_{X_H^{\mathcal{A}}} P = 0 = \mathcal{L}_{X_H^{\mathcal{A}}} H_i \text{ for all } H, \text{ and } \mathcal{L}_X P = 0 = \mathcal{L}_X H_i.$$

4.6.2 Nambu Mechanics and Measure-Constraints Preserving Flows

In the previous sections we have employed bi-vector fields to construct $\mathcal{A}$-mechanical systems, which associated to any Hamiltonian H a H-preserving vector field $X_H^{\mathcal{A}} \equiv \mathcal{A}^\sharp(\mathrm{d}H) = -i_{\mathrm{d}H}\mathcal{A}$. We have then noted that the class of bi-vector fields $\mathcal{A}$ that were P-preserving, that is $\mathrm{div}_P(X_H^{\mathcal{A}}) = 0$ for all H, were exactly those for which $\mathrm{curl}_P(\mathcal{A}) = 0$, or equivalently $\mathcal{A} = \mathrm{curl}_P(\mathcal{V})$ for some 3-vector field $\mathcal{V}$ (up to topological obstructions), so that our H and P-preserving flow was

$$X_H^{\mathcal{A}} = -i_{\mathrm{d}H}\mathrm{curl}_P(\mathcal{V}) = \mathrm{curl}_P(i_{\mathrm{d}H}\mathcal{V}).$$

In (4.48) we have built on this idea to construct systems that preserved both P and a set of constraints $\{H_i\}$ by setting $X \equiv \mathrm{curl}_P(i_{\mathrm{d}H_1\wedge\cdots\wedge\mathrm{d}H_\ell}\mathcal{V})$. Such vector fields are particular instances of *Nambu mechanics*, and we now show that these provide us with an elegant framework to understand the relation between differential geometry and measure-constraints preserving systems.

For $k \geq 1$, we define a k-**Nambu mechanics** to be a k-vector field $\mathcal{W}$, which we view as the set $\mathfrak{X}_{\mathcal{W}}(\mathcal{M})$ containing all the vector fields resulting from assigning to $k-1$ functions (or constraints) $\{H_i\}$ the vector field

$$X^{\mathcal{W}}_{H_1,\ldots,H_{k-1}} \equiv -i_{\mathrm{d}H_1\wedge\cdots\wedge\mathrm{d}H_{k-1}}\mathcal{W},$$

when $k \geq 2$, while we view a 1-Nambu system $X \in \mathfrak{X}(\mathcal{M})$ as the singleton $\mathfrak{X}_X(\mathcal{M}) \equiv \{X\}$, i.e., a map from 0 constraints (the empty set) to X, or $X^X_\emptyset \equiv X$. Note that 2-Nambu mechanics are what we previously referred to as $\mathcal{A}$-mechanics. It follows that we may think of the P-musical

isomorphisms as identifying $(n-k)$-twisted forms with k-Nambu mechanics for $k \geq 1$, while for $k = 0$ the P-musical isomorphism is simply the Radon-Nikodym derivative restricted to smooth measures. Moreover, we have the following key result:

Theorem 31. *We say a k-Nambu mechanics $\mathcal{W}$ is P-preserving iff*

$$div_P(X) = 0 \quad \forall X \in \mathfrak{X}_{\mathcal{W}}(\mathcal{M}),$$

which for $k \geq 2$ can also be written as

$$div_P(X^{\mathcal{W}}_{H_1,\dots,H_{k-1}}) = 0 \quad \forall H_j \in C^\infty(\mathcal{M}). \tag{4.49}$$

P-preserving k-Nambu mechanics are precisely those for which $curl_P(\mathcal{W}) = 0$. Hence, the P-musical isomorphism identifies closed $(n-k)$-twisted forms with P-preserving k-Nambu mechanics. Moreover, it identifies exact $(n-k)$-twisted forms with (P-preserving) k-Nambu mechanics arising from a globally curled k-vector field.

The proof is analogous to the derivation of theorem 11, and follows by recursively commuting curl_P with the Schouten–Nijenhuis bracket, so that the statement (4.49) becomes equivalent to the statement

$$i_{\mathrm{d}H_1\wedge\dots\wedge\mathrm{d}H_{k-1}}\mathrm{curl}_P(\mathcal{W}) = \mathrm{curl}_P(\mathcal{W})\,(\mathrm{d}H_1,\dots,\mathrm{d}H_{k-1}) = 0, \quad \forall H_j \in C^\infty(\mathcal{M}),$$

which implies $\mathrm{curl}_P(\mathcal{W}) = 0$ and thus, by theorem 3,

$$\mathcal{W} \in \mathrm{curl}_P\left(\mathfrak{X}^{k+1}(\mathcal{M})\right) \oplus P^\sharp\left(H^{n-k}\left(\mathcal{M}\right)\right).$$

The cases $k = 1, 2$ were the ones we used to characterise P-preserving vector fields (and thus P-preserving diffusions) and P-preserving $\mathcal{A}$-mechanics respectively; this theorem generalises those results to arbitrary k (the case $k = 0$ is not interesting since top twisted forms are always closed).

Having understood the intimate relation between geometry and measure-preserving flows, we can go even further. The above theorem characterises the properties that $\mathcal{W}$ must satisfy in order for *all* the elements of the Nambu-mechanics $X^{\mathcal{W}}_{H_1,\dots,H_k}$ to be P-divergence free. When $\mathcal{W}$ has vanishing rotationnel, and we fix a given set of independent constraints $f_1, \dots f_k$, the resulting flow is thus both P-preserving and $\{f_i\}$-preserving. This raises the important question of completeness, i.e., whether any P-preserving and $\{f_i\}$-preserving flow can be written as an element of a Nambu mechanics $X^{\mathcal{W}}_{f_1,\dots,f_k}$, for some appropriate $\mathcal{W}$. The answer is essentially yes: any such vector field is of the form of (4.48) up to a topological obstruction term which is now

$X = i_{\mathrm{d}H}\mathcal{A}$	$\Longrightarrow$	X preserves H
X is locally $\mathrm{curl}_P(\mathcal{A})$	$\Longleftrightarrow$	X preserves measure P
$\mathcal{B}$ antisymmetric	$\Longleftrightarrow$	X_f preserves associated density $\forall f$
$\mathcal{B}$ is locally $\mathrm{curl}_P(\mathcal{V})$	$\Longleftrightarrow$	X_f preserves measure P $\forall f$
$X = i_{\mathrm{d}H\wedge\mathrm{d}h}\mathcal{V}$	$\Longrightarrow$	X preserves both H and h
$X = \mathrm{curl}_P\left(i_{\mathrm{d}H\wedge\mathrm{d}h}\mathcal{W}\right)$	$\Longrightarrow$	X preserves both P, H and h
$\mathcal{V}$ is locally $\mathrm{curl}_P(\mathcal{W})$	$\Longleftrightarrow$	$X^{\mathcal{V}}_{f,h}$ preserves measure P $\forall f, h$
$[X_f, X_h] = X_g$, i.e., $(X_f, [\cdot,\cdot])$ Lie subalgebra	$\Longleftrightarrow$	Splitting Method has Shadow Hamiltonian
$X_{\{f,h\}} = [X_f, X_h]$ for all f, h	$\Longleftrightarrow$	$\{\cdot,\cdot\}$ satisfies left Leibniz identity
$X_{\{f,h\}} = [X_f, X_h]$ and $X_f(f) = 0$	$\Longleftrightarrow$	$\mathcal{B}$ Poisson bivector field
$\mathrm{div}_{\mu_\mathcal{M}}(X_{p_\infty} - p_\infty Y) = 0$	$\Longleftrightarrow$	$X_{-\log p_\infty} + Y$ preserves $p_\infty \mu_\mathcal{M}$
Poisson cohomology class of ϕ_P is trivial	$\Longleftrightarrow$	There is a Hamiltonian measure for Poisson Π

Table 4.2: A summary of some of the results .

associated to the **leafwise de Rham cohomology** induced by the foliation of $\mathcal{M}$ defined by the constraints.

Hence in this chapter, with no assumption other than smoothness, we have obtained a complete and canonical recipe of measure and constraints preserving flows on arbitrary manifolds. See table 4.2 for a summary of some of the results.

4.6.3 History of Measure-Preserving Diffusions

The history of measure-preserving processes is a long one, and we here only aim to provide a brief synopsis. While in the machine learning community the characterisation of measure-preserving diffusions on $\mathbb{R}^n$ was popularised by the recent NeurIPS article [242], analogous and much older results can be found in the mathematical literature. For example, in 1977 Robert Graham discussed the covariance of the Fokker–Planck equation in the context of non-degenerate diffusions, and used the Riemannian metric associated to the noise to define a Riemannian divergence which allowed him to differentiate second-order tensors [140] (a nice discussion of the work of Graham can be found in [112]). By analogy with Maxwell's equations, Graham noted that the Fokker–Planck current must be the Riemannian divergence of some anti-symmetric tensor field, which is precisely the result provided by theorem 2 in [242] (at least when the divergence with respect to the Lebesgue measure is interpreted as a Riemannian divergence). The covariance of the Fokker–Planck equation and the diffusion process is also discussed in [31, 256], and [256] derived conditions for the diffusion to be reversible. A less intuitive characterisation of the Fokker–Planck current of measure-preserving diffusions is also given in [171]. The book [176] effectively derives a complete recipe in the case of non-degenerate diffusions on compact orientable manifolds, using the Riemannian metric of the

noise to transform the Fokker–Planck vector field into a 1-form, that is then studied via its Hodge-de Rham decomposition. The main issues with the previous references is that they all assume the noise was non-degenerate, in order to equip the manifold with a Riemannian metric. Yet, many important measure-preserving diffusions are not in this form, in particular (the deterministic) Hamiltonian mechanics itself, and even the underdamped Langevin process. While on Euclidean space we have a "natural" metric, that we can use to differentiate second-order tensors, such metrics do not exist on manifolds, and, as we have shown, Riemannian metrics are irrelevant for characterising P-preserving diffusions. Recently (after the submission of this thesis) a new covariant formulation was introduced in [99] to remove the dependence on the metric in the work of Graham. Contrary to our recipe, the latter work relies on local coordinates, and moreover does not take into account the presence of topological obstructions, thus leading to a recipe which is only valid locally: on manifolds there are divergence-free vector fields that cannot be (globally) written as the P-rotationnel of a bivector field, for example those induced by non-zero harmonic forms on compact manifolds. Furthermore, the relation with the canonical geometry of P is not shown, yet this intrinsic geometry allows us to apply differential geometry theorems and go way beyond a characterisation of P-preserving diffusions as we have proved above. As another example pertinent to diffusions, recall that in §4.5 we were interested in generalising Langevin systems, in the sense of P-preserving diffusions with no volume terms. In that subsection we have focused on those with vertical noise, but note that it is straightforward to characterise the general class of P-preserving diffusions for which the volume terms vanish, $\mathrm{curl}_{\mu_\mathcal{M}}(\mathcal{A}) = 0 = \mathrm{div}_{\mu_\mathcal{M}}(Y_i)Y_i$: whenever the noise fields are (point-wise) linearly independent, then this statement is equivalent to $\mathcal{A} = \mathrm{curl}_{\mu_\mathcal{M}}(\mathcal{V})$ and $Y_i = \mathrm{curl}_{\mu_\mathcal{M}}(\mathcal{A}_i)$, or in other words the set of **volume-free P-preserving diffusions** is parametrised by 3-vector fields and (noise) 2-vector fields (or, more generally, by closed twisted $(n-3)$-forms and closed twisted $(n-2)$-forms when there are topological obstructions). Hence, any volume-free (in the sense above) $e^{-\beta H}\mu_\mathcal{M}$-preserving diffusion has the form (it is entirely made off curls)

$$\mathrm{d}Z_t = \mathrm{curl}_{\mu_\mathcal{M}}(i_{\mathrm{d}H}\mathcal{V})\,\mathrm{d}t + \tfrac{1}{2}\beta\mathrm{div}_{\mu_\mathcal{M}}(i_{\mathrm{d}H}\mathcal{A}_i)\mathrm{curl}_{\mu_\mathcal{M}}(\mathcal{A}_i)\mathrm{d}t + \mathrm{curl}_{\mu_\mathcal{M}}(\mathcal{A}_i)\circ\mathrm{d}W_t^i.$$

The fact that these volume-free diffusions are generated by higher-order multi-vector fields confer them with geometric properties that make them suitable candidates to generalise SOL-HMC.

In the bigger mathematical picture, we see that the characterisation of P-preserving diffusions is a combination of two things, the construction of the Fokker–Planck equation [300, 120, 193, 306], and its expression in geometric and divergence-form as discussed above (see also [287]), combined with representations of divergence-free vector field. Such representations have been known for at least a century [96]: for example the fact that vector fields that preserve the

Lebesgue measure can be written as the divergence of an antisymmetric matrix, which is all that is needed to obtain [242, Thrm. 2], goes back at least to the work of Poincaré and Volterra in the 1880s [355].

4.6.4 The Hamiltonian Conspiracy

In this chapter we have derived the characterisation of measure-constraints preserving flows, and provided straightforward recipes to construct them. Given how easy it is to build such flows, one may wonder how come many state-of-the-art samplers rely on Hamiltonian flows. For example, we have shown what if $\mathcal{A} = \mathrm{curl}_P(\mathcal{V})$, then the vector field $X_H^{\mathcal{A}}$ is P-preserving *for any* H, so finding a few functions H for which $X_H^{\mathcal{A}}$ is tractable seems like a feasible task. This is however not the case in general, because unless both $\mathcal{A}$ and $\mathrm{d}H$ have nice symmetries, $X_H^{\mathcal{A}}$ will lack symmetries and thus be typically intractable. Indeed, one of the initial motivations for the construction of the theory of Lie groups was precisely the analysis of conditions under which a flow could be explicitly integrated, leading to the concept of *solvable Lie algebras* [222, 65, 338]. In particular, the ability to integrate a flow by quadratures depends on the presence of appropriate symmetries and conserved quantities [71, Sec. 2]. Roughly, the idea is to use nice symmetries $Y \in \mathfrak{X}(\mathcal{M})$, i.e., $\mathcal{L}_Y X_H^{\mathcal{A}} = 0$, and conserved quantities f, i.e., $\mathrm{d}f(X_H^{\mathcal{A}}) = 0$, to simplify the equations of motion. The symmetry Y preserves the integral curves of $X_H^{\mathcal{A}}$, i.e., $(\Phi_t^Y)_* X_H^{\mathcal{A}} = X_H^{\mathcal{A}}$ (and thus the flow commute), which allows us to reduce the number of variables, while the conserved quantity f implies the flow remains tangent to the level sets $f =$ constant, thus reducing the dimension of the problem.

Example 32. *On $\mathcal{M} = \mathbb{R}^2$, write $X_H^{\mathcal{A}} = a(x,y)\partial_x + b(x,y)\partial_y$. Let $Y \equiv \partial_x$ and $f \equiv x$. Then $\mathcal{L}_Y X_H^{\mathcal{A}} = 0$ means the components are independent of x, i.e., $\partial_x a = \partial_x b = 0$, and thus $X_H^{\mathcal{A}} = a(y)\partial_x + b(y)\partial_y$. On the other hand, $\mathrm{d}f(X_H^{\mathcal{A}}) = 0$ means there is no motion along the x-axis, i.e., $a(x,y) = 0$, and thus $X_H^{\mathcal{A}} = b(x,y)\partial_y$.*

Example 33. *Let X be P-preserving, and $\mathcal{A}$ be any of its potentials, $X = curl_P(\mathcal{A})$. From §4.3.1.3 we see that f is a conserved quantity of X iff it is in the kernel of the modular operator $\phi_P(f) = 0$, i.e., $X_f^{\mathcal{A}}$ is P-preserving. Hence, the Hamiltonians of the P-preserving elements of the $\mathcal{A}$-mechanics are the conserved quantities of its curl vector field.*

In order to obtain a nice tractable flow we will thus need $X_H^{\mathcal{A}}$ to have symmetries, which is closely related to $\mathrm{d}H$ and $\mathcal{A}$ having symmetries, specifically

$$\mathcal{L}_Y X_H^{\mathcal{A}} = X_H^{\mathcal{L}_Y \mathcal{A}} + X_{Y(H)}^{\mathcal{A}},$$

but this generates a trade-off between the symmetries that we want $\mathcal{A}$ to have and its ability

to be P-preserving whenever P is a complex distribution:

Example 34. *Let $\mathcal{M} = \mathbb{R}^n$, and consider a target $P = e^{-H}\mathrm{d}x$ on $\mathbb{R}^n$. The simplest class of bi-vector fields on $\mathbb{R}^n$ are the constant ones, corresponding to translation-invariance, i.e., $\mathcal{A}$ can be identified with a constant anti-symmetric matrix. But such constant bi-vectors automatically satisfy $curl_{\mathrm{d}x}(\mathcal{A}) = 0$. Hence, since $curl_P(\mathcal{A}) = curl_{\mathrm{d}x}(\mathcal{A}) + X_H^{\mathcal{A}}$, the condition $curl_P(\mathcal{A}) = 0$ is equivalent to $X_H^{\mathcal{A}} = 0$, i.e., H is an $\mathcal{A}$-Casimir, a strenuous condition.*

This is precisely the reason why we expand the target measure P in terms of a nice measure $\mu_{\mathcal{M}}$, usually the invariant/uniform measure on $\mathcal{M}$, and construct brackets that preserve $\mu_{\mathcal{M}}$ and have nice symmetries. This may then be achieved using higher-order multi-vector fields that share the symmetries of $\mu_{\mathcal{M}}$ using the results of §4.6.1. Unfortunately, although this considerably simplifies the problem, even when $\mathcal{A}$ is very nice we may still not be able to implement $X_H^{\mathcal{A}}$. Indeed, as discussed in magnetic HMC [342, B.2] the constant symplectic matrix in non-canonical form

$$\mathcal{A} \equiv \begin{pmatrix} E & F \\ -F^\top & G \end{pmatrix}$$

does not yield tractable Hamiltonian flows even when generated by velocity-independent or position-independent Hamiltonians.

The combination of theoretical and practical guarantees is what makes Hamiltonian mechanics so powerful in the context of sampling. First, as we have discussed in §4.3, the class of mechanical systems that have both the shadow and volume-preserving properties is (essentially) given by unimodular Poisson manifolds. However these unimodular Poisson systems are particularly rare, aside from a few special cases, such as constant antisymmetric matrices on $\mathbb{R}^n$ (which are unimodular Poisson systems as shown above), they are not easy to build on general manifolds. The only unimodular Poisson system that can be easily constructed on arbitrary manifolds is precisely the canonical Hamiltonian mechanics on its cotangent bundle. Second, in most practical applications the target is expressed in terms of a (invariant) Riemannian measure, and as noted in §3.3.2, we can construct a Hamiltonian systems adapted to such P which combines a geodesic flow, and a vertical gradient flow, the latter of which is always tractable. Concerning the geodesic flow, as discussed in §3.8 and §3.9, we can often construct Riemannian metrics with sufficient symmetries to ensure the geodesic flow can be easily implemented on many homogeneous manifolds. For the other main class of manifolds, the level sets, the geodesic flow of the induced Riemannian metric is typically intractable, but in that case we can still employ discrete Lagrangian integrators, such as RATTLE with reversiblity check, to implement HMC without bias (recall §3.5.2), and these will still have a shadow Hamiltonian due to Hamilton–Jacobi theorem.

Chapter 5

Statistical Models, Reproducing Kernel Hilbert Space, Gradient Descent, and the Geometry of Stein Operators

In this short chapter we shall introduce some concepts that we will need for our final chapter §6. We shall also briefly mention some other work I have done during the Ph.D. [29, 62, 80], but decided not to discuss in detail to keep the focus on the geometry of smooth measures and brackets. We begin by introducing the notions of statistical models, statistical manifolds, and divergence functions in §5.1 since such concepts are often left ambiguous in the HMC literature, or their definitions are inconsistent with the ones used in information geometry, and this will enable us to formulate precisely Riemannian manifold HMC [131]. In §5.2 we will discuss Riemannian and natural gradients, as well as their descent schemes, which we will use in §6 to optimise the minimum Stein discrepancy estimators (and in §A.8 for minimum MMD estimators). Vector-valued reproducing kernel Hilbert spaces will be introduced in §5.3, and in §5.4 we will review Stein's method and the construction of Stein operator based on the generator of diffusions. Combining these ideas we will then define in §5.5 the Maximum Mean and Kernelised Stein Discrepancies (MMD and KSD). In the next chapter we shall also show that the bracket formalism allows us to easily extend the density construction of Stein operators beyond $\mathbb{R}$, and prove that the most general generator-based Stein operator is a particular instance of this more general construction.

5.1 Statistical Divergences and Information Metric

In §5.1.1 we will define the notion of statistical model. We view statistical models as sets of probability measure "embedded" in the space of signed measures, and use this embedding to build the geometry of statistical models, in particular the Fisher Riemannian metric. This is analogous to using $\mathbb{R}^n$ to derive geometric objects on embedded submanifolds. In §5.1.2 we define statistical (Banach) manifolds, which are statistical models parametrised by a Banach manifold. This yields an extrinsic construction of statistical manifolds which relates directly to probability measures, and is thus appropriate in the statistical context. An alternative construction of statistical manifolds relies on divergences which we introduce in §5.1.3.[1]

5.1.1 Statistical Models and the Fisher Riemannian Metric

A statistical model S is a subset of the set $\mathcal{P}(\mathcal{X})$ of probability measures $\mathcal{X}$. The set $\mathcal{P}(\mathcal{X})$ does not have a norm or vector space structure with respect to which we can differentiate functions. However it is a subset of the Banach space $\mathcal{S}(\mathcal{X})$ of finite signed measures on $\mathcal{X}$, equipped with the total variation norm $\|\cdot\|_{TV}$ (see (A.2)). This enables us to define tangent vectors to S: we say $v \in \mathcal{S}(\mathcal{X})$ is a vector tangent to $\mu \in \mathsf{S}$ if $v = \nabla|_0\gamma(1)$ for a Fréchet differentiable curve γ on S through $\mu = \gamma(0)$.[2] We then define the **tangent space** $T_\mu\mathsf{S}$ over $\mu \in \mathsf{S}$ as the linear space of tangent vectors to μ. The tangent space $T_\mu\mathsf{S}$ may be shown to be a subspace of the space $\mathcal{S}_0(\mathcal{X}, \mu)$ of signed measures dominated by μ which integrate to 0. Using the Banach space isomorphism $(\mathcal{S}(\mathcal{X}, \mu), \|\cdot\|_{TV}) \to \left(L^1(\mathcal{X}, \mu), \|\cdot\|_{L^1(\mathcal{X},\mu)}\right)$, $\nu \mapsto \frac{\mathrm{d}\nu}{\mathrm{d}\mu}$, we can further identify tangent vectors with μ-integrable functions (see 2 [22]):

$$T_\mu\mathsf{S} \subset \mathcal{S}_0(\mathcal{X}, \mu) \subset \mathcal{S}(\mathcal{X}, \mu) \cong L^1(\mathcal{X}, \mu).$$

Viewing tangent vectors $v \in T_\mu\mathsf{S}$ as μ-integrable functions, when these are further square-integrable, we may pull-back the inner product of the Hilbert space $L^2(\mathcal{X}, \mu)$ to $T_\mu\mathsf{S}$ in order to define the **Fisher Riemannian metric** g^F

$$g^F_\mu(u, v) \equiv \left\langle \frac{\mathrm{d}u}{\mathrm{d}\mu}, \frac{\mathrm{d}v}{\mathrm{d}\mu} \right\rangle_{L^2(\mathcal{X},\mu)} = \int_{\mathcal{X}} \frac{\mathrm{d}u}{\mathrm{d}\mu}\frac{\mathrm{d}v}{\mathrm{d}\mu}\mathrm{d}\mu.$$

[1] In the more mathematically inclined literature, finite dimensional statistical manifolds are instead often (equivalently) defined as finite dimensional manifolds equipped with a Riemannian metric and a totally-symmetric tensor of rank 3, which implies the the Riemannian manifold can be equipped with two torsion-free connections ∇ and ∇^* that are dual (or conjugate) to each other.

[2] To be precise $\gamma : (-T, T) \to S \subset S(\mathcal{X})$ where $\nabla|_0\gamma \in \mathrm{Hom}\,(\mathbb{R}, \mathcal{S}(\mathcal{X}))$ is the Fréchet derivative evaluated at 0, which we view as an element of $\mathcal{S}(\mathcal{X})$ by applying it to $1 \in \mathbb{R}$.

If all distributions in S are dominated by some fixed reference measure μ, these can be identified with their densities with respect to μ, so that S is viewed as a subset of $L^1(\mathcal{X}, \mu)$

$$\mathsf{S} \cong \{f \in L^1(\mathcal{X}, \mu) : f = \frac{\mathrm{d}Q}{\mathrm{d}\mu} \text{ with } Q \in \mathsf{S}\}.$$

If $t \mapsto Q(t) \in \mathsf{S}$ is a curve of distributions in S and $\gamma(t) \equiv \frac{\mathrm{d}Q(t)}{\mathrm{d}\mu}$, so $Q(t) = \gamma(t)\mu$, then if v is the tangent vector to Q at $Q(0)$, under the identification above we have (assuming μ and $Q(0)$ are equivalent measures)

$$\begin{aligned} v = \dot{Q}(0) = \dot{\gamma}(0)\mathrm{d}\mu \sim \frac{\mathrm{d}\dot{Q}(0)}{\mathrm{d}Q(0)} = \frac{\mathrm{d}\dot{Q}(0)}{\mathrm{d}\mu}\frac{\mathrm{d}\mu}{\mathrm{d}Q(0)} = \frac{\mathrm{d}\dot{Q}(0)}{\mathrm{d}\mu}\left(\frac{\mathrm{d}Q(0)}{\mathrm{d}\mu}\right)^{-1} = \frac{\dot{\gamma}(0)}{\gamma(0)} \\ = \frac{\mathrm{d}}{\mathrm{d}t}\log\gamma(t)\Big|_{t=0}. \end{aligned}$$

As a result, when viewed as functions tangent vectors are sometimes called **logarithmic tangent vectors**, and these are particularly natural in the context of finite dimensional parametrised models discussed next.

5.1.2 Parametrised Statistical Manifolds

The statistical models appearing in statistics are usually parametrised models. We say S is a **statistical manifold** if $\mathsf{S} = \phi(\Theta)$ where Θ is a Banach manifold and $\phi : \Theta \to \mathcal{S}(\mathcal{X})$ is differentiable. One of the most famous examples is the statistical manifold of Gaussian measures on $\mathcal{X} \equiv \mathbb{R}^n$, which is parametrised through the manifold $\Theta \equiv \mathbb{R}^n \times \mathbf{SPD}(n)$, where $\mathbf{SPD}(n)$ denotes the space of covariance matrices

$$\phi(m, \Sigma) \equiv \mathcal{N}(m, \Sigma) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} e^{-(x-m)^\top \Sigma^{-1}(x-m)}\mathrm{d}x. \tag{5.1}$$

It turns out that if Θ is finite dimensional then the statistical manifold $\phi(\Theta)$ is necessarily dominated by some measure [21], say ν. Let us write $\phi(\theta) \equiv p(\cdot|\theta)\mathrm{d}\nu \equiv p_\theta \mathrm{d}\nu$ for a **likelihood** function $p : \mathcal{X} \times \Theta \to \mathbb{R}$. [3] We can pullback the Fisher metric from S to Θ using ϕ, which defines a positive semi-definite symmetric bilinear form on Θ: if $t \mapsto \theta(t), t \mapsto \alpha(t)$ are any two curves on Θ with $\alpha(0) = \theta(0) = \beta$, then

$$(\phi^* g^F)_\beta \left(\dot{\theta}(0), \dot{\alpha}(0)\right) \equiv g^F_{\phi(\beta)}\left(\partial_\beta\phi\left(\dot{\theta}(0)\right), \partial_\beta\phi\left(\dot{\alpha}(0)\right)\right) = \left\langle \frac{\mathrm{d}\partial_\beta\phi\left(\dot{\theta}(0)\right)}{\mathrm{d}\phi(\beta)}, \frac{\mathrm{d}\partial_\beta\phi\left(\dot{\alpha}(0)\right)}{\mathrm{d}\phi(\beta)} \right\rangle_{L^2(\mathcal{X}, \phi(\beta))}$$

[3] Note this differs from the Bayesian notion of likelihood, which is just a function of the parameters.

$$= \left\langle \frac{\mathrm{d}\frac{\mathrm{d}\phi\circ\theta}{\mathrm{d}t}(0)}{\mathrm{d}\phi(\beta)}, \frac{\mathrm{d}\frac{\mathrm{d}\phi\circ\alpha}{\mathrm{d}t}(0)}{\mathrm{d}\phi(\beta)} \right\rangle_{L^2(\mathcal{X},\phi(\beta))} = \left\langle \frac{\mathrm{d}\frac{\mathrm{d}p_\theta}{\mathrm{d}t}(0)\nu}{\mathrm{d}\phi(\beta)}, \frac{\mathrm{d}\frac{\mathrm{d}p_\alpha}{\mathrm{d}t}(0)\nu}{\mathrm{d}\phi(\beta)} \right\rangle_{L^2(\mathcal{X},\phi(\beta))}$$
$$= \left\langle \frac{\mathrm{d}\log p_\theta}{\mathrm{d}t}(0), \frac{\mathrm{d}\log p_\alpha}{\mathrm{d}t}(0) \right\rangle_{L^2(\mathcal{X},\phi(\beta))} = \int_{\mathcal{X}} \frac{\mathrm{d}\log p_\theta}{\mathrm{d}t}\Big|_{t=0} \frac{\mathrm{d}\log p_\alpha}{\mathrm{d}t}\Big|_{t=0} p_\beta \mathrm{d}\nu.$$

From this derivation we note the identification $\partial_{\theta(0)}\phi(\dot{\theta}) \sim \frac{\mathrm{d}\log p_\theta}{\mathrm{d}t}\big|_{t=0}$. In particular if θ^i are local coordinates around β, we may apply the above to the coordinate curves and using $\frac{\mathrm{d}\log p_\theta}{\mathrm{d}t}\big|_{t=0} = \frac{\partial \log p_\theta}{\partial \theta^i}\big|_{\theta(0)} \frac{\mathrm{d}\theta^i}{\mathrm{d}t}\big|_{t=0}$ we obtain

$$(\phi^* g^F(\theta))_{ij} = \int_{\mathcal{X}} \frac{\partial \log p_\theta}{\partial \theta^i} \frac{\partial \log p_\theta}{\partial \theta^j} \mathrm{d}P_\theta,$$

which is the familiar **Fisher information matrix**. Often this is referred to as the Fisher metric, but $\phi^* g^F$ will only be a Riemannian metric (i.e., positive definite) when ϕ is an immersion. The function $\log p_\theta$ is usually called the **log-likelihood** and its derivative $\mathrm{d}_\theta \log p_\theta = \frac{\partial \log p_\theta}{\partial \theta^i}\mathrm{d}\theta^i$ is also called the score,[4] whose components (under the identification) are simply given by the pushforward of the coordinate basis $\partial\phi(\partial_{\theta^i}) \sim \mathrm{d}_\theta \log p_\theta(\partial_{\theta^i}) = \frac{\partial \log p_\theta}{\partial \theta^i}$.

Naturally we can define other Riemannian metrics on statistical manifolds for example [334] uses the parametrisation (5.1) on the space of Gaussian distributions on $\mathbb{R}^n$ with mean zero. Using the identification with $\mathbf{SPD}(n)$ it is shown that the Riemannian metric $\langle X, Y\rangle_\Sigma \equiv \mathrm{tr}\,(X\Sigma Y)$ induces the 2-Wasserstein distance. This is generalised in [81] (see proposition 7) which considers Wasserstein statitiscal manifold obtained from an immersion of a parameter set into Wasserstein space, and discusses the associated gradient flow. In §A.4.9 we briefly discuss the most famous statistical model: the exponential family.

5.1.3 Divergence Functions

A **divergence or contrast function** on a manifold Θ is a differentiable function $D : \Theta \times \Theta \to [0, \infty)$ s.t., (see 4.4.1 [21])

$$\textbf{(i)} \quad D(x\|y) = 0 \text{ iff } x = y, \qquad \text{and} \qquad \textbf{(ii)} \quad (X, 0)(0, Y)D|_{(x,x)} < 0 \tag{5.2}$$

for any vector fields that do not vanish at x. Here (X, Y) denotes the vector field $X \oplus Y$ on $\Theta \times \Theta$ for vector fields $X, Y \in \mathfrak{X}(\Theta)$.[5] The second condition is not always included, but

[4] Note this is not the same "score" we considered §4, since we now differentiate over parameters.
[5] This is sometimes written as $X_x Y_y D(x\|y)|_{x=y} \equiv (X, 0)(0, Y)D|_{x=y}$

guarantees the "information" tensor

$$g^D(X,Y) \equiv -(X,0)(0,Y)D|_{x=y}$$

is a Riemannian metric on Θ. Locally

$$g^D(\theta)_{ij} = -\partial_{\theta^i}\partial_{\alpha^j} D(\theta\|\alpha)|_{\alpha=\theta}.$$

If only **(i)** holds we will say D is a **statistical divergence**. One of the most famous examples is the **Kullback–Leibler (statistical) divergence** given by

$$\mathrm{KL}(P\|Q) \equiv \int_{\mathcal{X}} \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}P$$

whenever the RHS is defined, and satisfies **(i)** of (5.2). On a parametrised statistical manifold dominated by ν we can pullback this expression to

$$D(\theta\|\alpha) \equiv \phi^* \,\mathrm{KL}(\theta\|\alpha) = \int_{\mathcal{X}} \log\left(\frac{p_\theta}{p_\alpha}\right) p_\theta \mathrm{d}\nu$$

whose associated bilinear form g^D is simply the Fisher information matrix. However KL is hard to estimate in the case of unnormalised statistical models (with unknown normalisation constants), so in §6 we will focus on score-based divergences, such as the **score-matching statistical divergence** defined for continuously differentiable positive densities on an open set $\mathcal{X} \subset \mathbb{R}^d$ as

$$\mathrm{SM}(P\|Q) \equiv \tfrac{1}{2}\int_{\mathcal{X}} p(x)\|\nabla_x \log p - \nabla_x \log q\|_2^2 \mathrm{d}x.$$

A final example that is a genuine divergence on $\mathbb{R}^n$ is the **Bregman divergence** of a strictly convex function $f : \mathbb{R}^n \to \mathbb{R}$

$$D_f(x,y) \equiv f(x) - f(y) - \nabla f(y)(x-y),$$

whose information tensor is the Hessian $g^{D_f} = \nabla^2 f$. There is a nice connection between discrete Lagrangians and information divergences as discussed in §A.5.6.5.

5.1.4 HMC on Statistical Manifolds

Riemannian manifold HMC (RMHMC) [131] received a lot of attention in statistics, and motivated many statisticians to look more deeply into the geometry of HMC. It also caused considerable confusion as: (i) despite its name the method samples from distributions on Euclidean

space; (ii) the original HMC method was in fact already a "Riemannian manifold HMC", since it samples from the gauge group SU(3). The original aim of RMHMC was to sample a posterior distribution $P \equiv P_{post}(\cdot|\omega) \in \mathcal{P}(\Theta)$ on parameter space $\mathcal{M} \equiv \Theta \equiv \mathbb{R}^n$, given some observation $\omega \in \mathcal{X}$ and a prior distribution $P_{prior} \in \mathcal{P}(\Theta)$. The idea is to choose a Riemannian metric on Θ which locally matches the Hessian of the posterior (see [237] for a discussion),

$$\Sigma_{post,\omega}(\theta) \equiv -\frac{\partial^2}{\partial\theta^i\partial\theta^j}\log\rho_{post}(\theta|\omega) = -\frac{\partial^2}{\partial\theta^i\partial\theta^j}\left(\log\rho(\omega|\theta) + \log\rho_{prior}(\theta)\right).$$

Taking expectations with respect to the conditional distribution $\propto \rho(\omega|\theta)\mathrm{d}\omega$ in order to get a ω-independent quantity yields (assuming the "integration by parts" is valid)

$$\begin{aligned}
\int \Sigma_{post,\omega}(\theta)\rho(\omega|\theta)\mathrm{d}\omega &= -\int \rho(\omega|\theta)\frac{\partial^2}{\partial\theta^i\partial\theta^j}\log\rho(\omega|\theta)\mathrm{d}\omega - \frac{\partial^2}{\partial\theta^i\partial\theta^j}\log\rho_{prior}(\theta) \\
&= \int \frac{\partial}{\partial\theta^i}\rho(\omega|\theta)\frac{\partial}{\partial\theta^j}\log\rho(\omega|\theta)\mathrm{d}\omega - \frac{\partial^2}{\partial\theta^i\partial\theta^j}\log\rho_{prior}(\theta) \\
&= \int \rho(\omega|\theta)\frac{\partial}{\partial\theta^i}\log\rho(\omega|\theta)\frac{\partial}{\partial\theta^j}\log\rho(\omega|\theta)\mathrm{d}\omega + \Sigma_{prior}(\theta) \\
&= \left(\phi^* g^F(\theta)\right)_{ij} + \Sigma_{prior}(\theta),
\end{aligned}$$

where in the last equation we recognise the Hessian of the prior, and the pullback of the Fisher metric defined on the statistical model of conditional measures $\mathsf{S} \equiv \{\rho(\cdot|\theta)\mathrm{d}\omega : \theta \in \Theta\}$, parametrised by $\phi : \Theta \to \mathsf{S}$, $\theta \mapsto \rho(\cdot|\theta)\mathrm{d}\omega$. Thus, setting $G(\theta) \equiv \phi^* g^F(\theta) + \Sigma_{prior}$, the target Hamiltonian of RMHMC is then (recall §2.1.2)

$$H(\theta, p) = -\log\rho_{post}(\theta|\omega) + \tfrac{1}{2}\log\det G(\theta) + \tfrac{1}{2}p^\top G(\theta)^{-1}p.$$

Since the geodesic flow of G is usually intractable, the standard leapfrog integrator cannot be used, but one can resort to using the generalised leapfrog integrator (see §A.5.5.1), with fixed-point iterations to deal with the implicit equations. A computationally cheaper alternative is discussed in [88] which applies the explicit integrator of [335] and compares it with the generalised leapfrog integrator on some statistical problems.

5.2 Natural Gradient Descent

Consider local coordinates q^i on a manifold $\mathcal{M}$, and a function $f \in C^\infty(\mathcal{M})$. The tuple $(\partial_1 f, \ldots, \partial_n f)$ is often called the gradient of f in statistics. In the context of optimisation the problem of estimating the minima of the "loss" function f is often tackled by trying to find

a curve that follows the direction of steepest descent of f in order to move efficiently towards the local minimum of f. The curve is defined in this coordinate system as the solution of the differential equation

$$\frac{\mathrm{d}q^i}{\mathrm{d}t} = -\frac{\partial f}{\partial q^i}.$$

Of course this equation is geometrically meaningless (even when $\mathcal{M} = \mathbb{R}^n$), since the LHS corresponds to components of the tangent vector field to the curve, while the RHS are the components of the differential of f, a 1-form. That is if $\gamma : [0, T] \to \mathcal{M}$ were the solution to this equation, and $\dot{\gamma}$ its tangent vector field (which is a section of the pullback bundle $\gamma^*(T\mathcal{M})$ over $[0, T]$), then the equation reads $\dot{\gamma} = -\mathrm{d}f$, which cannot hold since the two sides of this equation transform differently under change of coordinates (see also 1.4b [122]). It follows that to construct a geometric equation we must replace the RHS by some appropriate vector field associated to the function f. The answer to the problem "how do we construct such a vector field" becomes clear once we have recalled the problem we were trying to solve in the first place: find the direction of steepest descent. Indeed to give a meaning to the word "steepest" we must define a notion of steepness of the tangent vector, that is a norm on the tangent spaces. Such a norm is determined by first defining a Riemannian metric $G : \mathcal{M} \to T^*\mathcal{M} \otimes T^*\mathcal{M}$, which in turn defines a norm

$$\|\cdot\| \equiv \sqrt{G(\cdot,\cdot)} : T\mathcal{M} \to \mathbb{R}$$

on each tangent space. As we observed in §4.2.2 the metric maps any function f to a vector field ∇f called the **Riemannian gradient**. The **gradient flow** of f is then defined as the curve γ that solves the equation (i.e., the integral curve of the G^{-1}-Hamiltonian field)

$$\dot{\gamma} = -\nabla_\gamma f.$$

Let us check this definition is appropriate. Suppose γ is the gradient flow of f. That it represents direction of descent follows by considering the rate of change $\dot{\gamma}(f)$ of f along γ:

$$\dot{\gamma}(f) = \mathrm{d}f(\dot{\gamma}) = \mathrm{d}f\left(-\nabla f\right) = -\mathrm{d}f(\nabla f) = -G(\nabla f, \nabla f) = -\|\nabla f\|^2 \leq 0.$$

That it is the direction of steepest descent follows by observing that the gradient direction yields the dual norm of $\mathrm{d}f$,

$$\|\mathrm{d}_p f\| \equiv \sup_{v \in T_p\mathcal{M} : \|v\| = 1} \left|\mathrm{d}_p f(v)\right|,$$

since by Cauchy–Schwarz inequality $\left|\mathrm{d}_p f(v)\right| = \left|-G\left(\nabla_p f, v\right)\right| \leq \|\nabla_p f\|\|v\|$, hence $\|\mathrm{d}_p f\| \leq \|\nabla_p f\|$ and the upper bound is attained at $v = \nabla_p f / \|\nabla_p f\|$ (i.e., the normalised gradient).

The above discussion also holds when $\mathcal{M} \equiv \mathcal{H}$ is a Hilbert space. In that case the exterior derivative becomes the Fréchet derivative $\mathrm{d}f : \mathcal{H} \to \mathcal{H}^*$, which has an associated vector field

$\nabla f : \mathcal{H} \to \mathcal{H}$ by Riesz representation theorem. The gradient flow γ is thus defined just as above

$$\dot{\gamma} = -\nabla f(\gamma), \quad \text{or equivalently} \quad \langle \dot{\gamma}, v \rangle_{\mathcal{H}} = -\langle \nabla f(\gamma), v \rangle_{\mathcal{H}} \quad \forall v \in \mathcal{H}.$$

For example, **Stein variational gradient descent** (SVGD) is constructed using the gradient of the functional $f : \mathcal{H}^d \to \mathbb{R}$ on the vector RKHS $\mathcal{H}^d$ on $\mathbb{R}^d$ with kernel $K = k\mathrm{Id}$ (see §5.3)

$$f(h) \equiv \mathrm{KL}\left(q_{\mathrm{Id}+h} \| p\right)$$

where p, and $q_{\mathrm{Id}+h}$ are the densities with respect to the Lebesgue measure of P and $(\mathrm{Id}+h)_\sharp P$ respectively. The SVGD algorithm updates its current approximation $q = q_{\mathrm{Id}}$ of the target P by following the gradient of f at $h = 0$ to move along a "perturbation" of q that best decreases the KL divergence with the target. From theorem 3.3 [230] this direction is given by

$$\nabla_h \,\mathrm{KL}\left(q_{\mathrm{Id}+h} \| p\right)|_{h=0} = -\int \frac{1}{p(x)} \nabla_x \left(p k_x\right) \mathrm{d}\mathbb{Q}(x).$$

While this formula assumes $\mathbb{Q}$ has a density q, SVGD uses discrete measure $\frac{1}{\ell}\sum_{i=1}^{\ell} \delta_{x_i}$, associated to a set of "particles" $\{x_i\}_i^\ell$, to approximate the target using the RHS: given a step size $\eta > 0$ the discrete measure is updated by updating its particles with

$$\begin{aligned} x_i &\leftarrow x_i + \frac{\eta}{\ell} \sum_j \frac{1}{p(x_j)} \nabla_x \left(p k_x\right)|_{x_j}(x_i) \\ &= x_i + \frac{\eta}{\ell} \sum_j \left(k(x_j, x_i) \nabla_{x_j} \log p + \nabla_1 k(x_j, x_i)\right) \equiv x_i + \eta \Delta_p x_i. \end{aligned}$$

A particular instance of descent method, based on the Fisher information metric, was developed by Amari and collaborators [12]. It is a widely used alternative to Euclidean gradient descent methods. However as noted in §5.1.3 the Fisher information matrix is not necessarily positive definite since KL is only a statistical divergence (rather than a divergence), and as a result we refer to the descent method associated with KL, or a more general statistical divergence, as **natural gradient descent**, rather than Riemannian gradient descent. Natural gradient descent has been successfully applied to a variety of problems in machine learning and statistics, for example reinforcement learning [181], neural network training [290], Bayesian variational inference methods [158] and Markov chain Monte Carlo [131]. While the classical natural gradient approach is based on the Fisher information matrix induced by the KL divergence, information geometries arising from other metrics on probabilities have also been studied in previous works, including those arising from optimal transport metrics [81] and the Fisher divergence [184].

Let us now discuss the practical implementation of the gradient descent $\dot{\gamma}(t) = -\nabla_{\gamma(t)} f$ on

a complete connected Riemannian manifold. Typically [6] the gradient flow is approximated by the update equation $\gamma(t+1) = \exp_{\gamma(t)}(-\eta_t H(Z_t, \gamma))$ where exp is the Riemannian exponential map, (η_t) is a sequence of step sizes with $\sum \eta_t^2 < \infty$, $\sum \eta_t = +\infty$, and H is an unbiased estimator of the loss gradient, $\mathbb{E}[H(Z_t, \gamma)] = \nabla_\gamma f$. When the Riemannian exponential is computationally expensive, it is convenient to replace it by a retraction $\mathcal{R}$ (see §A.5.2). This leads to the update $\gamma(t+1) = \mathcal{R}_{\gamma(t)}(-\eta_t H(Z_t, \gamma))$ [51]. On a linear manifold it is common to take $\mathcal{R}_{\gamma(t)}(-\eta_t H(Z_t, \gamma)) \equiv \gamma(t) - \eta_t H(Z_t, \gamma(t))$, which is the update we will use in §6. Therein we will be interested in learning a probability distribution $\mathbb{Q}$, by following the gradient flow of a statistical divergence $f(\cdot) \equiv D(\mathbb{Q}\|\cdot)$ over a statistical model S. Given a parametrisation of S with local coordinates (θ^i), we can locally pullback the gradient method to the parametrisation. We thus consider the parametrisation $t \mapsto \theta(t)$ of the curve γ and locally expand the natural gradient in terms of the information tensor

$$\nabla_\theta D(\mathbb{Q}\|\theta) = g^D(\theta)^{-1} \partial_\theta D(\mathbb{Q}\|\theta),$$

where $\partial_\theta D$ denotes the tuple $(\partial_{\theta^i} D)$. We will approximate this natural gradient using the biased estimator

$$H(\{X_i^t\}_i, \theta) \equiv \hat{g}_{\theta(t)}(\{X_i^t\}_{i=1}^n)^{-1} \partial_\theta \hat{D}(\{X_i^t\}_{i=1}^n \| \theta),$$

where $\hat{g}_{\theta(t)}(\{X_i^t\}_{i=1}^n)$ and $D(\{X_i^t\}_{i=1}^n\|\theta)$ are unbiased estimators for the information matrix $g^D(\theta(t))$ and $D(\mathbb{Q}\|\theta)$ using a sample $\{X_i^t\}_{i=1}^n \sim \mathbb{Q}$. We thus obtain the following natural gradient descent algorithm

$$\theta(t+1) = \theta(t) - \eta_t \hat{g}_{\theta(t)}(\{X_i^t\}_{i=1}^n)^{-1} \partial_{\theta(t)} \hat{D}(\{X_i^t\}_{i=1}^n \| \theta). \tag{5.3}$$

Note, when $\eta_t = \frac{1}{t}$, g^D is the Fisher metric, and $\hat{D}(\{X_i^t\}_{i=1}^n\|\theta)$ is replaced by the estimator of the KL-statistical divergence $\widehat{\mathrm{KL}}(\{X_i^t\}_{i=1}^n\|\theta)$ obtained by replacing the $\mathbb{Q}$-integral of the likelihood by a U-statistic, we recover the natural gradient descent algorithm [12]. The statistical divergences we employ in §6 will often be constructed from vector-valued reproducing kernels, which we now introduce.

5.3 Reproducing Kernel Hilbert Spaces

Given a Hilbert space $\mathcal{W}$, a Hilbert space $\mathcal{H}$ of maps $\mathcal{X} \to \mathcal{W}$ is called a **reproducing kernel Hilbert space** (RKHS) if the evaluation maps $\delta_x : \mathcal{H} \to \mathcal{W}$, $\delta_x f = f(x)$, are continuous for all

[6] See also sec 4.4 [124] for Riemannian Newton method

$x \in \mathcal{X}$. We define its **(reproducing) kernel** $K : \mathcal{X} \times \mathcal{X} \to B(\mathcal{W})$, where $B(\mathcal{W})$ is the space of bounded linear operators on $\mathcal{W}$ [293],

$$K(x, y) = \delta_x \circ \delta_y^*.$$

Note $K(x, y)^* = K(y, x)$, and since δ_x is bounded, its adjoint $\delta_x^* : \mathcal{W} \to \mathcal{H}$ is also a bounded linear map usually denoted K_x. We thus have

$$\langle K_x g, f\rangle_{\mathcal{H}} \equiv \langle \delta_x^* g, f\rangle_{\mathcal{H}} = \langle g, \delta_x f\rangle_{\mathcal{W}} = \langle g, f(x)\rangle_{\mathcal{W}}\,,$$

and

$$K(x, y)w = \delta_x \circ \delta_y^* w = \left(\delta_y^* w\right)(x) = \left(K_y w\right)(x),$$

so

$$\langle g, K(x, y)w\rangle_{\mathcal{W}} = \langle g, \left(K_y w\right)(x)\rangle_{\mathcal{W}} = \langle \delta_x^* g, K_y w\rangle_{\mathcal{W}} = \langle K_x g, K_y w\rangle_{\mathcal{W}}\,.$$

It follows that for any $x_1, \ldots, x_N \in \mathcal{X}$ the operator $(K(x_i, x_j))$ on $\mathcal{W}^N$ is positive semidefinite, since for any $w_1, \ldots, w_N \in \mathcal{W}$

$$\langle K(x_i, x_j)w_j, w_i\rangle_{\mathcal{W}} = \left\langle K_{x_i} w_i, K_{x_j} w_j\right\rangle_{\mathcal{W}} = \|K_{x_i} w_i\|_{\mathcal{W}}^2 \geq 0.$$

Moreover the set of functions of the form $K_{x_j} w_j$ is dense in $\mathcal{H}$. Conversely, by theorem 2.1 [262], given a positive semidefinite function $K : \mathcal{X} \times \mathcal{X} \to B(\mathcal{W})$ satisfying $K(x, y)^* = K(y, x)$, there exists up to isometry a unique RKHS of maps $\mathcal{X} \to \mathcal{W}$ whose reproducing kernel is K. A **feature map** is a map $\phi : \mathcal{X} \to B(\mathcal{G}, \mathcal{W})$, where $\mathcal{G}$ is a Hilbert space called the **feature space**, which defines a reproducing kernel by

$$K(x, y) \equiv \phi(x) \circ \phi(y)^* \in B(\mathcal{W}),$$

and elements of the induced RKHS $\mathcal{H}$ may expressed as $f(\cdot) = \phi(\cdot)g$ for some $g \in \mathcal{G}$ [262]. The map $\delta : x \mapsto \delta_x$ is an example of feature map. When $\mathcal{W} = \mathbb{R}$, $\phi : \mathcal{X} \to B(\mathcal{G}, \mathbb{R}) = \mathcal{G}^*$, and the dual $\phi(x)^* \in \mathcal{G}$ is simply the dual vector associated to the linear functional $\phi(x)$ by Riesz representation theorem. Since $B(\mathbb{R}) = \mathbb{R}^* \cong \mathbb{R}$, the above equation takes the standard form

$$K(x, y) = \langle \phi(x)^*, \phi(y)^*\rangle_{\mathcal{G}}\,. \tag{5.4}$$

For our purpose the set $\mathcal{X}$ will be an open subset of $\mathbb{R}^d$, and $\mathcal{W}$ will be either $\mathbb{R}^d$ or $\mathbb{R}$. The case $\mathcal{W} = \mathbb{R}^d$ (with standard inner product) corresponds to **vector-valued RKHS**, where the kernel $K : \mathcal{X} \times \mathcal{X} \to B(\mathbb{R}^d) \cong \mathbb{R}^{d\times d}$ may be viewed as a matrix-valued kernel with components $(K(x, y))_{ji} \equiv e_j \cdot K(x, y)e_i$ (where (e_i) is the canonical basis of $\mathbb{R}^d$), satisfying

$K(x,y)^T = K(y,x)$. Theorem 2.11 [263] summarises the relation between continuity and smoothness of K, and the associated continuity and smoothness of functions in $\mathcal{H}$. When $\mathcal{W} = \mathbb{R}$, K is called a **scalar reproducing kernel**, and is usually denoted by $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Chapter 7 [36] provides many examples of scalar reproducing kernels and their RKHS. Given a scalar kernel k, an important example of vector-valued RKHS is given by defining the product of Hilbert spaces $\mathcal{H}^d$ with inner product $\langle f, g \rangle_{\mathcal{H}^d} \equiv \sum_i \langle f_i, g_i \rangle_{\mathcal{H}}$. This is a vector-valued RKHS with matrix kernel $K(x,y) = k(x,y)\mathrm{Id}$.

Kernels are ubiquitous in machine learning and statistics. Kernel algorithms often arise through the **kernel trick**: if an algorithm depends only on inner products $\langle x, y \rangle$, replace these by $k(x,y)$ to obtain its "kernelised" version. In this manner one obtains non-linear generalisations of the algorithm, from the assumption that while the algorithm shouldn't depend linearly in x, y, it should be possible to find a feature map s.t., the algorithm is linear in some higher dimensional feature space. Indeed from (5.4) we see that kernel trick replaces $\langle x, y \rangle$ by $\langle \phi^*(x), \phi^*(y) \rangle$. See [142] for a discussion of the construction of kernelised algorithms for PCA and regression. A useful result in optimisation is the **representer theorem**, which states that if K is a matrix kernel on $\mathbb{R}^d$, and $R : (\mathbb{R}^d)^\ell \times \mathbb{R}_{\geq 0} \to \mathbb{R}$ is any function strictly increasing in its second entry, then any local minimiser of the functional $J : \mathcal{H} \to \mathbb{R}$, $J(f) \equiv R\left((f(x_i))_{i=1}^\ell, \|f\|_{\mathcal{H}}^2\right)$, takes the form (theorem 1 [261]) $\sum_i^\ell K_{x_i} w_i$, for some $w_i \in \mathbb{R}^d$. For example the family of functionals $J(f) \equiv \sum_i L(y_i, f(x_i)) + \eta \|f\|_{\mathcal{H}}^2$, for some $y_i \in \mathbb{R}^d$, where $L : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, satisfy the assumption of the theorem, and appear in kernel ridge regression. There is also a bijection between the covariance functions of Gaussian processes and scalar kernels. A discussion of the connection between Gaussian processes and kernel ridge regression can be found in [182]. See also [82, 159, 359] for kernel methods in deep learning, regression, classification, and approximation.

5.4 Stein Operators from Generators of Diffusions

We now describe an important application of generators to construct Stein operators which will be used in §6 to define minimum Stein discrepancy estimators, though we will later show these may obtained more naturally from geometric considerations. We begin by briefly reviewing Stein's method §5.4.1. We then consider the Langevin and diffusion Stein operators in §5.4.2, which will be generalised in §6.9, by showing they are special cases of the generalisation of the modular vector field of Weinstein.

5.4.1 Stein's Method

The aim of Stein's method is to bound the error between a distribution P on $\mathcal{X}$ of interest, originally a Gaussian [327], and other distributions, in order to quantify the quality of the approximation, and prove convergence in distribution to P. The idea is as follows. We first consider an integral probability metric [274]

$$d_{\mathcal{H}}(P,Q) \equiv \sup_{f\in\mathcal{H}} \left| \int f \mathrm{d}P - \int f \mathrm{d}Q \right| \tag{5.5}$$

over an appropriate family $\mathcal{H}$ of functions $\mathcal{X} \to \mathbb{R}$. In statistics this is often written as a uniform bound on the expectations

$$d_{\mathcal{H}}(X,W) \equiv \sup_{f\in\mathcal{H}} |\mathbb{E}[f(X)] - \mathbb{E}[f(W)]|$$

for random variables $X \sim P, W \sim Q$. Examples of integral probability metrics include the 1-Wasserstein, Total variation, and Kolmogorov metrics. We refer to [130] for a review of metrics on probability measures and the relations between them. Given P, the method relies on finding a **Stein operator** $\mathcal{T}$ (often denoted $\mathcal{S}_P$) whose domain is a **Stein class** of functions $\mathcal{F}(\mathcal{T})$ that characterises the target P in the sense that

$$\int \mathcal{T}[f]\mathrm{d}Q = 0 \quad \forall f \in \mathcal{F}(\mathcal{T}), \qquad \text{iff} \qquad Q = P. \tag{5.6}$$

Here it is assumed that the range $\mathcal{T}\left(\mathcal{F}(\mathcal{T})\right)$ consists of $\mathbb{R}$-valued functions that are integrable over an appropriate set $\{Q\}$ of distributions of interest. In the original case $P = \mathcal{N}(0,1)$, the Stein operator

$$\mathcal{T}[f] = \frac{\mathrm{d}f}{\mathrm{d}x} - f \cdot Id$$

characterises $\mathcal{N}(0,1)$ for a suitable Stein class as shown in Stein's Lemma [309, lemma 2.1]. Given such a pair $(\mathcal{T}, \mathcal{F}(\mathcal{T}))$ we can then formulate **Stein equation** for $f_h \in \mathcal{F}(\mathcal{T})$ given $h \in \mathcal{H}$:

$$\mathcal{T}[f_h] = h - \int h \mathrm{d}P.$$

Integrating over Q and taking the supremum over $\mathcal{H}$ yields

$$d_{\mathcal{H}}(Q,P) = \sup_{f\in\mathcal{F}(\mathcal{H})} \left| \int \mathcal{T}[f]\mathrm{d}Q \right|$$

where $\mathcal{F}(\mathcal{H})$ is the set of solutions to Stein equation for $h \in \mathcal{H}$, and the aim of Stein's method is then to bound the RHS using properties of Q and $\mathcal{T}$ (see [217, 218])

5.4.2 Diffusion Stein Operator

Techniques from Stein's method were also proposed to **measure sample quality**, that is to quantify the discrepancy between the target P and a discrete distribution $Q = w^i \delta_{q^i}$ where $\sum_i w^i = 1$. The aim is to obtain a discrepancy which is computationally tractable and controls convergence in $d_{\mathcal{H}}$. Specifically for $\mathcal{X} \equiv \mathbb{R}^n$, [137] consider **Stein discrepancies** defined as

$$\mathrm{SD}(Q,P) \equiv d_{\mathcal{T}[\mathcal{H}]}(Q,P) = \sup_{h\in\mathcal{H}} \left| \int \mathcal{T}[h] \mathrm{d}Q \right| = \sup_{h\in\mathcal{H}} \left| w^i \mathcal{T}[h](q^i) \right|$$

where $\mathcal{H}$ is a set of functions $\mathbb{R}^n \to \mathbb{R}^n$, and $\mathcal{T}$ an operator on $\mathcal{H}$ satisfying $\int \mathcal{T}[h]\mathrm{d}P = 0$ for all $h \in \mathcal{H}$. The construction of $\mathcal{T}$ therein relies on the observation that the generator $\mathcal{L}$ of the diffusion (4.1) with target P satisfies

$$\int \mathcal{L}[f]\mathrm{d}P = 0$$

under appropriate conditions on $\mathcal{L}$ and f, which is called the **generator approach** [24]. For a target $P = \rho \mathrm{d}q$, the Smoluchowsky dynamics $\mathrm{d}Z_t = \nabla \log \rho(Z_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$ (see §4.1.2) has generator

$$\mathcal{L} = \nabla \log \rho \cdot \partial_q + \Delta,$$

which led [137] to the **Langevin Stein operator**

$$\mathcal{T}[h] \equiv \nabla \log \rho \cdot h + \nabla \cdot h$$

defined over a set $\mathcal{H}$ given in proposition 1 therein. In [136] the authors proposed a more general family of operators called **diffusion Stein operators**, which may be obtained from (4.5) by ignoring the antisymmetric component and formally replacing $\frac{1}{2}\Sigma\nabla h$ with mh where m is an arbitrary $n \times n$ matrix

$$\mathcal{S}[h] \equiv \frac{1}{\rho} \nabla \cdot (\rho m h)\,.$$

We can also extract it more directly from the generator by rewriting it as (using (4.6))

$$\begin{aligned}
\mathcal{L}f &= b \cdot \nabla f + \tfrac{1}{2} \operatorname{tr}\left(\Sigma \nabla^2 f\right) = \tfrac{1}{2}\rho^{-1}\nabla \cdot (\rho(\Sigma + A)) \cdot \nabla f + \tfrac{1}{2} \operatorname{tr}\left(\Sigma \nabla^2 f\right) \\
&= \tfrac{1}{2}\rho^{-1}\nabla \cdot (\rho\Sigma) \cdot \nabla f + \tfrac{1}{2}\rho^{-1} \operatorname{tr}\left(\rho\Sigma \nabla^2 f\right) + \tfrac{1}{2}\rho^{-1}\nabla \cdot (\rho A) \cdot \nabla f \\
&= \tfrac{1}{2}\rho^{-1}\nabla \cdot (\rho\Sigma\nabla f) + \tfrac{1}{2}\rho^{-1}\nabla \cdot (\rho A) \cdot \nabla f = \tfrac{1}{2}\rho^{-1}\nabla \cdot (\rho\Sigma\nabla f) + \tfrac{1}{2}\rho^{-1}\nabla \cdot (\rho A \nabla f) \\
&= \tfrac{1}{2}\rho^{-1}\nabla \cdot (\rho(\Sigma + A)\nabla f)\,,
\end{aligned}$$

where we have used §A.1 and $\mathrm{tr}(AS) = 0$ for A, S anti-symmetric and symmetric matrices respectively. The fact that $\nabla^2 f$ is a symmetric matrix is precisely the reason we need to first write the generator in this form before making the substitution $\nabla f \mapsto h$. Replacing $\frac{1}{2}(\Sigma + A)$ with m and ∇f with h yields the diffusion Stein operator.

Note the Stein discrepancy defined by the Stein diffusion operators only requires knowledge of ρ up to normalisation, and does not involve any integral over P (these are typically intractable). On the other hand it does require solving an optimisation problem over $\mathcal{H}$ which is usually infinite dimensional. One way to circumvent this issue is to choose $\mathcal{H}$ to be the unit ball of a reproducing kernel Hilbert space as demonstrated, which for the Langevin Stein operator leads to the kernel Stein discrepancy as we discuss next.

5.5 Representers and Kernel Stein Discrepancy

Consider a scalar reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with RKHS $\mathcal{H}$. If $\mathbb{P} \in \mathcal{P}(\mathcal{X})$ is a distributions on $\mathcal{X}$, recall the $\mathcal{H}$-valued map $x \mapsto k_x$ is Bochner $\mathbb{P}$-integrable iff $x \mapsto \|k_x\|_{\mathcal{H}} = \sqrt{k(x,x)}$ is $\mathbb{P}$-integrable (see A.5.4 [328]). If $\mathcal{P}_k(\mathcal{X})$ denotes the set of distributions s.t., $x \mapsto k_x$ is Bochner integrable, we define the **mean embedding**

$$\mu : \mathcal{P}_k(\mathcal{X}) \to \mathcal{H}, \qquad \mu(\mathbb{P}) \equiv \mu_{\mathbb{P}} \equiv \int k_x \mathrm{d}\mathbb{P}(x).$$

By Proposition 2 [326] $\mathcal{P}_k(\mathcal{X}) = \mathcal{P}(\mathcal{X})$ iff k is bounded. Moreover by Lemma 3.1 [273], $\mu(\mathbb{P})$ represents the $\mathbb{P}$-expectation functional on $\mathcal{H}$:

$$\int f \mathrm{d}\mathbb{P} = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}, \qquad \forall f \in \mathcal{H}.$$

The mean embedding $\mathcal{P}_k(\mathcal{X}) \to \mathcal{H}$ allows us to transfer structures from $\mathcal{H}$ to $\mathcal{P}_k(\mathcal{X})$. For example, we can define a pseudometric on $\mathcal{P}_k(\mathcal{X})$, called the **maximum mean discrepancy** (MMD), as the pullback by μ of the Hilbert space metric $d^{\mathcal{H}}$ induced by the inner product

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) \equiv \left(\mu^* d^{\mathcal{H}}\right)(\mathbb{P}, \mathbb{Q}) = d^{\mathcal{H}}(\mu_{\mathbb{P}}, \mu_{\mathbb{Q}}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}.$$

Hence MMD is a metric on $\mathcal{P}_k(\mathcal{X})$ iff μ is injective, in which case we say the kernel is **characteristic**. Theorems 7 and 9 [326] determine conditions under which k is characteristic (though these can be easily improved). The name of MMD results from the fact that using the representer property of μ, it is easy to show MMD is the integral pseudometric (5.5) associated to the unit ball $B_1(\mathcal{H})$, and thus corresponds to the maximal difference of means over $B_1(\mathcal{H})$ (or

equivalently the worst-case integration error over $B_1(\mathcal{H})$): if $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_k(\mathcal{X})$

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) = d_{B_1(\mathcal{H})}(\mathbb{P}, \mathbb{Q}) = \sup_{\|f\|_{\mathcal{H}} \leq 1} \left| \int f \mathrm{d}\mathbb{P} - \int f \mathrm{d}\mathbb{Q} \right| = \sup_{\|f\|_{\mathcal{H}} \leq 1} |\langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}}| = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}.$$

Moreover using Bochner integrability and A.32 [328] (and continuity of the linear functional $\langle h, \cdot \rangle_{\mathcal{H}}$ for any $h \in \mathcal{H}$) the squared MMD (which is a statistical divergence) has the particularly simple expression

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q})^2 = \int k \mathrm{d}\mathbb{P} \otimes \mathrm{d}\mathbb{P} + \int k \mathrm{d}\mathbb{Q} \otimes \mathrm{d}\mathbb{Q} - 2 \int k \mathrm{d}\mathbb{P} \otimes \mathrm{d}\mathbb{Q}. \tag{5.7}$$

From the discussion above it follows that if k is bounded and characteristic MMD is a metric on $\mathcal{P}(\mathcal{X})$, and theorem 21 [326] provides bounds between MMD and Dudley, Wasserstein and total variation distances, while theorems 23-24 yield settings where MMD metrises narrow convergences on $\mathcal{P}(\mathcal{X})$. MMD and the mean embedding have many applications, including two sample tests [143], training generative neural networks [107], MCMC and inference [273], and is related to **energy distances** which include the "distance covariance", a discrepancy that measures the dependency of random variables (corollary 26 [312]). Computing the MMD^2 statistical divergence between $\mathbb{P}$ and $\mathbb{Q}$ only requires integrating the kernel with respect to their product distributions. In particular given samples from $\mathbb{P}$ and $\mathbb{Q}$ we can use U-statistics as unbiased estimators of the divergence. In [62] we have carefully analysed the application of minimum MMD^2 estimators to density-free inference, for statistical generative models using natural gradient descent (see §A.8 for a summary of our results). Specifically, we have derived generalisation bounds, proved consistency and asymptotically normality of the estimators, and shown these were more robust to model misspecification than the competing Wasserstein-based estimators (the proofs of these results were done by Andrew Duncan and I).

If we know the density of $\mathbb{P}$ up to a normalising constant, we can use Stein operators to remove any integral in $\mathbb{P}$. Let $\mathcal{T}$ be a Stein operator as in §5.4.1 for $\mathbb{P}$ with Stein class $\mathcal{F}(\mathcal{T})$. In fact we only need that $\mathcal{T}$ satisfies the weaker requirement

$$\int \mathcal{T}[f] \mathrm{d}\mathbb{P} = 0 \quad \forall f \in \mathcal{F}(\mathcal{T}). \tag{5.8}$$

In statistics it is common to call an operator $\mathcal{T}$ and class of functions $\mathcal{F}(\mathcal{T})$ for which (5.8) holds a Stein operator and Stein class for $\mathbb{P}$ even though they do not characterise $\mathbb{P}$ in the sense of (5.6). The **kernel Stein discrepancy** (KSD) associated with k is defined by considering

the squared IPM of the image under $\mathcal{T}$ of the vector RKHS $\mathcal{H}^d$ defined by $K \equiv k\mathrm{Id}$ (see §5.3)

$$\mathrm{KSD}(\mathbb{P}, \mathbb{Q}) \equiv d_{\mathcal{T}B_1(\mathcal{H}^d)}(\mathbb{P}, \mathbb{Q})^2 = \sup_{f \in \mathcal{H}^d : \|f\|_{\mathcal{H}^d} \leq 1} \left| \int \mathcal{T} f \mathrm{d}\mathbb{Q} \right|^2 .$$

For $\mathcal{X} = \mathbb{R}^d$ a standard choice is the Langevin Stein operator (see §5.4.2). Then (5.8) becomes essentially a divergence theorem

$$\int_{\mathbb{R}^d} \nabla \cdot (pf) \mathrm{d}x = 0, \quad \forall f \in \mathcal{F}(\mathcal{T})$$

and holds for example under conditions given by proposition 1 [138]. By theorems 3.6 and 3.8 [229], if k, p are sufficiently smooth and $k_x \in \mathcal{F}(\mathcal{T})$ for any x, then we can handle the supremum and obtain

$$\mathrm{KSD}(\mathbb{P}, \mathbb{Q}) = \int k_0 \mathrm{d}\mathbb{Q} \otimes \mathrm{d}\mathbb{Q}$$

where $k_0 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is the (Langevin) **Stein kernel**

$$k_0(x, y) \equiv k(x, y) \nabla_x \log p \cdot \nabla_y \log p + \nabla_x \log p \cdot \nabla_y k_x + \nabla_y \log p \cdot \nabla_x k_y + \nabla_x \cdot \nabla_y k_x.$$

Following the proof of theorem 1 [283] we have

$$k_0(x, y) = \left\langle \frac{1}{p(x)} \nabla_x (p k_x), \frac{1}{p(y)} \nabla_y (p k_y) \right\rangle_{\mathcal{H}^d} .$$

Comparing with (5.4) we deduce that k_0 is the (scalar) reproducing kernel (hence the name) of the Hilbert space of functions

$$\mathcal{H}_0 = \{\phi(\cdot) f = \langle \phi(\cdot)^*, f \rangle_{\mathcal{H}^d} : \mathcal{X} \to \mathbb{R} : f \in \mathcal{H}^d\},$$

where we have used the feature map $\phi : \mathcal{X} \to (\mathcal{H}^d)^*$, $\phi(x)^* \equiv \frac{1}{p(x)} \nabla_x (p k_x)$. Moreover, noting that $\int k_0(x, \cdot) \mathrm{d}\mathbb{P}(x) = 0$ and comparing with (5.7), we observe that $\mathrm{MMD}_{k_0}(\mathbb{P}, \mathbb{Q})^2$ built on the RKHS $\mathcal{H}_0$ is precisely the $\mathrm{KSD}(\mathbb{P}, \mathbb{Q})$ defined above

$$\mathrm{KSD}(\mathbb{P}, \mathbb{Q}) = \mathrm{MMD}_{k_0}(\mathbb{P}, \mathbb{Q})^2.$$

It is worth mentioning that unlike MMD^2, KSD is not symmetric since the Stein operator $\mathcal{T}$ depends on its entries. Indeed, labelling Stein operators with the distribution they characterise, we have

$$\mathrm{KSD}(\mathbb{P}, \mathbb{Q}) = d_{\mathcal{T}_{\mathbb{P}} B_1(\mathcal{H}^d)}(\mathbb{P}, \mathbb{Q})^2 \neq d_{\mathcal{T}_{\mathbb{Q}} B_1(\mathcal{H}^d)}(\mathbb{Q}, \mathbb{P})^2 = \mathrm{KSD}(\mathbb{Q}, \mathbb{P}).$$

However it provides a discrepancy measure between distributions that only depends on the

unnormalised density of its first entry and an integration over its second entry. It can become a statistical divergence under additional assumptions on k as given for example in proposition 3.3 [229]. Moreover given independent $\mathbb{Q}$-distributed random variables (X_i), the U-statistic

$$\frac{1}{n(n-1)} \sum_{i<j} k_0(X_i, X_j)$$

is an unbiased estimator of $\mathrm{KSD}(\mathbb{P}, \mathbb{Q})$, while the V-statistic is the KSD between $\mathbb{P}$ and the (random) empirical distribution $\mathbb{Q}_n$ of $\mathbb{Q}$

$$\mathrm{KSD}(\mathbb{P}, \mathbb{Q}_n) = \frac{1}{n^2} \sum_{ij} k_0(X_i, X_j).$$

KSD has been successfully applied to measure sample quality [138], control variates for Monte Carlo integration on Euclidean space [283], goodness-of-fit tests and model evaluation [229]. In [80] we have proposed several methods to approximate distributions with discrete measures based on minimising the V-statistics of KSD above. The algorithms therein substantially reduce the computational cost of implementing the original Stein point method (which relied on an intractable optimisation process) [79], by using local updates constructed using reversible MCMC transition kernels. In particular, we have shown that the method was consistent and controlled narrow convergence when MALA transitions were employed (under mild assumptions), with substantial empirical improvements against competing methods, such as SVGD, for Bayesian inference (the proofs were done by Lester Mackey and I).

Chapter 6

Minimum Stein Discrepancy Estimators for Unnormalised Models

Maximum likelihood estimation (see §2.2.1) is a de facto standard for estimating the unknown parameters in a statistical model $\{\mathbb{P}_\theta : \theta \in \Theta\}$. However, the computation and optimisation of a likelihood typically requires access to the normalising constants of the model distributions. This poses difficulties for complex statistical models for which direct computation of the normalisation constant would entail prohibitive multidimensional integration of an unnormalised density. Examples of such models arise naturally in modelling images [146, 196], natural language [267], Markov random fields [310] and nonparametric density estimation [324, 360]. To by-pass this issue, various approaches have been proposed to address parametric inference for unnormalised models as we now discuss. In the reminder we shall denote the target distribution as $\mathbb{Q}$ instead of P, to emphasise that we no longer consider smooth target distributions: the target is now only known through some samples it generates.

In **Markov Chain Monte Carlo maximum likelihood** [129], to implement MLE, that is

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \log p_\theta(X)$$

where X is some data and p_θ the joint density (recall (2.5)) is only known up to normalising constant, so we know $\overline{p}_\theta$ where $p_\theta = \frac{1}{Z(\theta)}\overline{p}_\theta$, we choose an arbitrary positive density f (with respect to the same reference measure "dx") known up to normalisation, $f = \frac{1}{Z}\overline{f}$, and equivalently

consider maximising the likelihood ratio

$$\hat{\theta} = \arg\max_{\theta\in\Theta} \log \frac{p_\theta(X)}{f(X)} = \arg\max_{\theta\in\Theta} \left(\log \frac{\overline{p}_\theta(X)}{\overline{f}(X)} - \log \frac{Z(\theta)}{Z} \right).$$

The final term on the RHS can be rewritten as an expectation

$$\frac{Z(\theta)}{Z} = \frac{\int \overline{p}_\theta(x)\mathrm{d}x}{Z} = \frac{\int \frac{\overline{p}_\theta}{\overline{f}}(x)\overline{f}(x)\mathrm{d}x}{Z} = \int \frac{\overline{p}_\theta}{\overline{f}}(x) f(x)\mathrm{d}x.$$

Given m independent $f\mathrm{d}x$-distributed random variables (Y_i), we can then build an estimator for the MLE,

$$\hat{\theta}_m \equiv \arg\max_{\theta\in\Theta} \left(\log \frac{\overline{p}_\theta(X)}{\overline{f}(X)} - \log \left(\frac{1}{m} \sum_{j=1}^{m} \frac{\overline{p}_\theta(Y_j)}{\overline{f}(Y_j)} \right) \right).$$

See [129, 128] for details about the convergence of this estimator to MLE.

When $\mathcal{X}$ is a finite set, $\Theta = \mathbb{R}^m$, and we write the discrete model distributions as $\mathbb{P}_\theta \equiv \frac{1}{Z_\theta} e^{-E_\theta}$, as we discussed in §5.2, a standard approach to locally maximise a function is by gradient ascent. However the (random) Euclidean gradient of the likelihood at θ

$$\begin{aligned}
\frac{\partial \frac{1}{n}\sum_i \log \mathbb{P}_\theta(X_i)}{\partial\theta} &= -\frac{1}{n}\sum_i \frac{\partial E_\theta(X_i)}{\partial\theta} - \frac{\partial \log Z_\theta}{\partial\theta} = -\frac{1}{n}\sum_i \frac{\partial E_\theta(X_i)}{\partial\theta} - \frac{\partial \log \sum_x e^{-E_\theta(x)}}{\partial\theta} \\
&= -\frac{1}{n}\sum_i \frac{\partial E_\theta(X_i)}{\partial\theta} + \frac{\sum_x e^{-E_\theta(x)} \partial_\theta E_\theta(x)}{\sum_y e^{-E_\theta(y)}} = -\mathbb{Q}_n\left(\frac{\partial E_\theta}{\partial\theta}\right) + \mathbb{P}_\theta\left(\frac{\partial E_\theta}{\partial\theta}\right)
\end{aligned}$$

requires computing an intractable expectation over $\mathbb{P}_\theta$ (here again we denote by $\mathbb{Q}_n$ the empirical distribution of the target $\mathbb{Q}$). To circumvent this issue [153] proposed instead to follow the gradient flow of the **contrastive divergence**

$$\mathrm{CD}_m \equiv \mathrm{KL}(\mathbb{Q}_n \| \mathbb{P}_\theta) - \mathrm{KL}(\mathbb{Q}_n \mathbb{P}^m \| \mathbb{P}_\theta) \tag{6.1}$$

where $\mathbb{P}^m$ is the m-step transition kernel of an ergodic $\mathbb{P}_\theta$-invariant Markov chain (see §A.4.10), and $\mathbb{Q}_n\mathbb{P}^m$ means we have averaged the initial value of the chain with respect to $\mathbb{Q}_n$. When $m = \infty$ the second term vanishes and CD_∞ reduces to KL. On the other hand for finite m we have

$$\begin{aligned}
\frac{\partial \mathrm{CD}_m}{\partial\theta} &= \frac{\partial}{\partial\theta}\sum_x \mathbb{Q}_n(x) \log \frac{\mathbb{Q}_n(x)}{\mathbb{P}_\theta(x)} - \frac{\partial}{\partial\theta}\sum_x \mathbb{Q}_n\mathbb{P}^m(x) \log \frac{\mathbb{Q}_n\mathbb{P}^m(x)}{\mathbb{P}_\theta(x)} \\
&= \mathbb{Q}_n\left(\frac{\partial E_\theta}{\partial\theta}\right) - \mathbb{Q}_n\mathbb{P}^m\left(\frac{\partial E_\theta}{\partial\theta}\right),
\end{aligned}$$

that is we have replaced the intractable expectation with an expectation over the distribution

of the Markov chain at time m. Unfortunately, the estimators will in general be biased (see 2 [153]). See also noise-contrastive estimation [76, 145, 146]. A **Stein contrastive divergence** learning was proposed in [228]. Here $\mathcal{X} = \mathbb{R}^d$, and, inspired by the CD_m objective function, the authors propose to replace the Markov chain transport of $\mathbb{Q}_n$ by an SVGD perturbation of the empirical distribution, leading to the objective

$$\theta \mapsto \Delta_{p_\theta} \mathbb{Q}_n \left(\log p_\theta\right) - \mathbb{Q}_n \left(\log p_\theta\right),$$

where $\Delta_{p_\theta} \mathbb{Q}_n$ is the (random) discrete measure with particles $X_i + \eta \Delta_{p_\theta} X_i$ (recall the notation from §5.2).

A related approach is the **minimum probability flow learning** [320]. Here the idea is to replace the problem

$$\operatorname*{argmin}_\theta \mathrm{KL}(\mathbb{Q}_n \| \mathbb{P}_\theta)$$

with

$$\operatorname*{argmin}_\theta \mathrm{KL}(\mathbb{Q}_n \| \mathbb{P}^\epsilon_\theta)$$

where $\mathbb{P}^\epsilon_\theta$ is the distribution at time ϵ of a continuous time $\mathbb{P}_\theta$-reversible Markov chain. Such a chain may be described in terms of its generator matrix Γ (see 2.1 [294]), which in [320] is set to

$$\Gamma_{ij} \equiv g_{ij} e^{\frac{1}{2}(E_\theta(i) - E_\theta(j))} \quad i \neq j, \qquad \Gamma_{ii} \equiv -\sum_{i \neq j} \Gamma_{ji}$$

for a suitable "connectivity" symmetric matrix (g_{ij}) describing which states can exchange probability. For small enough ϵ, $\mathrm{KL}(\mathbb{Q}_n \| \mathbb{P}^\epsilon_\theta)$ and its Euclidean gradient can be computed without knowledge of the normalisation constants of the model (see (14-16)).

The approaches discussed above often provide state-of-the-art estimations in practice, though the most used estimators still seem to be the ones based on the score matching statistical divergence. In §6.1 we will review in more detail the score matching estimator, its generalisations, and weaknesses. We will then introduce the class of minimum Stein discrepancy estimators §6.2, which only depend on the target distribution through expectations. Choosing the diffusion Stein operator further ensures these estimators only depend on the model through the gradients of the log-densities; thus making the estimators widely applicable. We will provide some important examples in §6.3, §6.4, §6.5, which generalise KSD and SM. In §6.6 we will derive the information metrics of the discrepancies, which may be used to implement natural gradient descent, and analyse the properties of the estimators in §6.7. In §6.8 we include the numerical experiments of [27]. Finally, in §6.9 and §6.10 we will discuss some more geometric aspects of Stein operators. Specifically, we will extend the "density approach", which constructs Stein operators on $\mathbb{R}$, to arbitrary manifolds, by showing that the Lie derivative provides a

canonical Stein operator (on vector fields). When combined with brackets this further shows that the "generator approach", which is the method of reference (except on $\mathbb{R}$), is in fact a special case of this generalised method and there is no need to introduce diffusions. We will then use the bracket formalism and geometry of Stein operators to derive a further generalisation of diffusion score-matching that may be used on manifolds. The remaining sections will be devoted to the derivation of some additional properties of the estimators.

6.1 Score Matching

The **score matching** (SM) estimator [173, 174] is a minimum score estimator [94] based on the Hyvärinen scoring rule (see §2.2.1) that avoids normalizing constants by depending on $\mathbb{P}_\theta$ only through the gradient of its log density $\nabla_x \log p_\theta$. It was originally proposed as a minimum score matching estimator, with the score matching defined as (recall §5.1.3)

$$\mathrm{SM}(\theta) \equiv \mathrm{SM}(\mathbb{Q}\|\mathbb{P}_\theta) = \tfrac{1}{2}\int_{\mathbb{R}^d} q(x)\|\nabla_x \log p_\theta - \nabla_x \log q\|_2^2 \mathrm{d}x.$$

SM estimators have proven to be a widely applicable method for estimation for models with unnormalised smooth positive densities. The above expression assumes the log density to be differentiable over the whole sample space $\mathbb{R}^d$, but generalisations to binary and non-negative data [174] as well as closed oriented Riemannian manifolds [246] have been produced. For the case of non-negative data, $\mathcal{X} \equiv (\mathbb{R}_{>0})^d$, one replaces the objective function by

$$\mathrm{SM}_{NN}(\theta) \equiv \tfrac{1}{2}\int_{(\mathbb{R}_{>0})^d} q(x)\|\nabla_x \log p_\theta \odot x - \nabla_x \log q \odot x\|_2^2 \mathrm{d}x,$$

where $\odot$ is the Hadamard product, and the terms "$\nabla_x f \odot x$" arise by differentiating the "scaling parametrisation" at the identity:

$$\nabla_\sigma f \circ s_x|_{\sigma=1} = \nabla_x f \odot x$$

where $s_x : (\mathbb{R}_{>0})^d \to (\mathbb{R}_{>0})^d$, $s_x(\sigma) \equiv x \odot \sigma$, since

$$\nabla_1 (f \circ s_x) = \nabla_x f \circ \nabla_1 s_x = \nabla_x f \cdot \mathrm{diag}(x_1, \ldots, x_d) = \nabla_x f \odot x.$$

See also [240] for a generalisation of the score matching objective that is more suitable to estimation of discrete measures. When $\mathcal{X}$ is an oriented Riemannian manifold with Riemannian volume form $\mathrm{vol}_\mathcal{X}$, we simply replace the derivative ∇f with the Riemannian gradient $\nabla^\mathcal{X} f$, to

obtain the objective function

$$\mathrm{SM}(\theta) \equiv \tfrac{1}{2} \int_{\mathcal{X}} q(x) \left\langle \nabla_x^{\mathcal{X}} \log p_\theta - \nabla_x^{\mathcal{X}} \log q, \nabla_x^{\mathcal{X}} \log p_\theta - \nabla_x^{\mathcal{X}} \log q \right\rangle_x^{\mathcal{X}} \mathrm{vol}_{\mathcal{X}}(\mathrm{d}x).$$

Expanding the inner product

$$\begin{aligned}
&\int_{\mathcal{X}} q(x) \left\langle \nabla_x^{\mathcal{X}} \log p_\theta - \nabla_x^{\mathcal{X}} \log q, \nabla_x^{\mathcal{X}} \log p_\theta - \nabla_x^{\mathcal{X}} \log q \right\rangle_x^{\mathcal{X}} \mathrm{vol}_{\mathcal{X}}(\mathrm{d}x) \\
&= \int_{\mathcal{X}} q \left(\|\nabla^{\mathcal{X}} \log p_\theta\|^2 - 2 \left\langle \nabla^{\mathcal{X}} \log p_\theta, \nabla^{\mathcal{X}} \log q \right\rangle^{\mathcal{X}} + \|\nabla^{\mathcal{X}} \log q\|^2 \right) \mathrm{vol}_{\mathcal{X}},
\end{aligned}$$

and by Stokes theorem

$$\begin{aligned}
&- \int_{\mathcal{X}} q \left\langle \nabla^{\mathcal{X}} \log p_\theta, \nabla^{\mathcal{X}} \log q \right\rangle^{\mathcal{X}} \mathrm{vol}_{\mathcal{X}} = - \int_{\mathcal{X}} \left\langle \nabla^{\mathcal{X}} \log p_\theta, \nabla^{\mathcal{X}} q \right\rangle^{\mathcal{X}} \mathrm{vol}_{\mathcal{X}} \\
&= \int_{\mathcal{X}} q \Delta \log p_\theta \mathrm{vol}_{\mathcal{X}} - \int_{\partial \mathcal{X}} q \left\langle \nabla^{\mathcal{X}} p_\theta, n \right\rangle^{\mathcal{X}} \mathrm{vol}_{\partial \mathcal{X}}.
\end{aligned}$$

Thus if the boundary term vanishes (for example $\mathcal{X}$ is a closed manifold) we obtain

$$\mathrm{SM}(\theta) = \tfrac{1}{2} \int_{\mathcal{X}} \left(\|\nabla^{\mathcal{X}} \log p_\theta\|^2 + 2 \Delta \log p_\theta \right) \mathrm{d}\mathbb{Q} + \tfrac{1}{2} \int_{\mathcal{X}} \|\nabla^{\mathcal{X}} \log q\|^2 \mathrm{d}\mathbb{Q}. \tag{6.2}$$

Since the last integral does not depend on θ it does not affect the minimiser of the objective, while the first integral, which involves a Riemannian version of the Hyvärinen score (2.9), may be estimated using a sample from $\mathbb{Q}$, and only requires knowing the statistical model up to normalising constant (we will generalise this in §6.10).

Despite the flexibility of this approach, SM has three important and distinct limitations. Firstly, as the Hyvärinen score depends on the Laplacian of the log-density, SM estimation will be expensive in high dimension and will break down for non-smooth models or for models in which the second derivative grows very rapidly. Secondly, as we shall demonstrate, SM estimators can behave poorly for models with heavy tailed distributions. Thirdly, the SM estimator is not robust to outliers in many applications of interest. Each of these situations arise naturally for energy models, particularly product-of-experts models and ICA models [175]. In a separate strand of research, new approaches have been developed to measure discrepancy between an unnormalised distribution and a sample based on Stein discrepancies, in particular KSD, that control weak convergence of an empirical measure to a target [134, 133, 245, 135].

In what follows we consider minimum Stein discrepancy (SD) estimators and show that SM, minimum probability flow and contrastive divergence estimators are all special cases. Within this class we focus on SDs constructed from RKHS, establishing the consistency, asymptotic normality and robustness of these estimators. Moreover, we shall demonstrate that these SDs

are appropriate for estimation of non-smooth distributions and heavy tailed distributions.

6.2 Minimum Stein Discrepancy Estimators

For our purposes we are interested in discrepancies that can be evaluated when $\mathbb{P}_\theta$ is only known up to normalisation, precluding the use of KL divergence. We instead construct estimators using Stein discrepancies based on Stein operators which do not depend on the normalisation. Let $\Gamma(\mathcal{Y}) \equiv \Gamma(\mathcal{X}, \mathcal{Y}) \equiv \{f : \mathcal{X} \to \mathcal{Y}\}$. As explained in (5.8) we weaken the requirements introduced in §5.4.1 and call a map $\mathcal{S}_\mathbb{P} : \mathcal{G} \subset \Gamma(\mathbb{R}^d) \to \Gamma(\mathbb{R})$ a Stein operator over a Stein class $\mathcal{G} \equiv \mathcal{F}(\mathcal{S}_\mathbb{P})$ if $\int_\mathcal{X} \mathcal{S}_\mathbb{P}[f] \mathrm{d}\mathbb{P} = 0 \; \forall f \in \mathcal{F}(\mathcal{S}_\mathbb{P})$ for any $\mathbb{P}$. From §5.4.2 the Stein discrepancy is then

$$\mathrm{SD}_{\mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]}(\mathbb{Q} \| \mathbb{P}_\theta) \equiv \sup_{f \in \mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]} \left| \int_\mathcal{X} f \mathrm{d}\mathbb{P}_\theta - \int_\mathcal{X} f \mathrm{d}\mathbb{Q} \right| = \sup_{g \in \mathcal{F}} \left| \int_\mathcal{X} \mathcal{S}_{\mathbb{P}_\theta}[g] \mathrm{d}\mathbb{Q} \right|. \tag{6.3}$$

The Stein discrepancy depends on $\mathbb{Q}$ only through expectations, and does not require the existence of a density, therefore permitting $\mathbb{Q}$ to be an empirical measure. If $\mathbb{P}$ has a C^1 density p on $\mathcal{X}$, recall from §5.4.2 that to obtain discrepancies which do not depend on the normalising constant of p, one can consider the Langevin-Stein discrepancy arising from the Stein operator $\mathcal{T}_p[g] \equiv \langle \nabla \log p, g \rangle + \nabla \cdot g$ [134, 133], or more generally, for an arbitrary $m \in \Gamma(\mathbb{R}^{d \times d})$ which we call **diffusion matrix**, the diffusion Stein operators: $\mathcal{S}_p^m[g] \equiv (1/p) \nabla \cdot (pmg)$ where $g \in \Gamma(\mathbb{R}^d)$, which we extend to matrix fields $\mathcal{S}_p^m[A] \equiv (1/p) \nabla \cdot (pmA)$, $A \in \Gamma(\mathbb{R}^{d \times d})$. The associated **minimum Stein discrepancy estimates** are then the minimisers of (6.3). As we will only have access to a sample $\{X_i\}_{i=1}^n \sim \mathbb{Q}$, we will focus on the **minimum Stein discrepancy estimators** minimising an approximation $\widehat{\mathrm{SD}}_{\mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]}(\{X_i\}_{i=1}^n \| \mathbb{P}_\theta)$ based on a U-statistic of the $\mathbb{Q}$-integral:

$$\hat{\theta}_n^{\mathrm{Stein}} \equiv \mathrm{argmin}_{\theta \in \Theta} \widehat{\mathrm{SD}}_{\mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]}(\{X_i\}_i^n \| \mathbb{P}_\theta). \tag{6.4}$$

Related and complementary approaches to inference using SDs include the nonparametric estimator of [219], the density ratio approach of [231] and the variational inference algorithms of [241, 304]. We now highlight several instances of SDs which will be studied in detail.

6.3 Example 1: Diffusion Kernel Stein Discrepancy Estimators

A convenient choice of Stein class is the unit ball of a RKHS of a scalar kernel function k, since this yields a closed-form expression for the supremum (see §5.5) For the Langevin Stein operator $\mathcal{T}_p$, the resulting KSD first appeared in [282] and has since been considered extensively in the context of hypothesis testing, measuring sample quality and approximation of probability measures in [79, 78, 83, 98, 135, 227, 226, 223]. We now discuss its application to inference, and will in fact consider the more general class of discrepancies based on the diffusion Stein operator and matrix-valued kernels for additional flexibility.

Consider an RKHS $\mathcal{H}^d$ of functions $f \in \Gamma(\mathbb{R}^d)$ with (matrix-valued) kernel $K \in \Gamma(\mathcal{X} \times \mathcal{X}, \mathbb{R}^{d\times d})$ (see §5.3), and we set $K_x \equiv K(x, \cdot)$. The Stein operator $\mathcal{S}_p^m[f]$ induces an operator $\mathcal{S}_p^{m,2}\mathcal{S}_p^{m,1} : \Gamma(\mathcal{X} \times \mathcal{X}, \mathbb{R}^{d\times d}) \to \Gamma(\mathbb{R})$ which acts first on the first variable and then on the second one. We briefly mention two simple examples of matrix kernels constructed from scalar kernels. If we want the components of f to be orthogonal, we can use the diagonal kernel (i) $K = \mathrm{diag}(\lambda_1 k^1, \dots, \lambda_d k^d)$ where $\lambda_i > 0$ and k^i is a C^2 kernel on $\mathcal{X}$, for $i = 1, \dots, d$; else we can "correlate" the components by setting (ii) $K = Bk$ where k is a (scalar) kernel on $\mathcal{X}$ and B is a (constant) symmetric positive definite matrix. We propose to study **diffusion kernel Stein discrepancies** indexed by K and m (see §6.11 for the proof):

Theorem 35 (**Diffusion Kernel Stein Discrepancy**). *For any kernel K, we find that $\mathcal{S}_p^m[f](x) = \langle \mathcal{S}_p^{m,1} K_x, f\rangle_{\mathcal{H}^d}$ for any $f \in \mathcal{H}^d$. Moreover if $x \mapsto \|\mathcal{S}_p^{m,1} K_x\|_{\mathcal{H}^d} \in L^1(\mathbb{Q})$, we have*

$$\mathrm{DKSD}_{K,m}(\mathbb{Q}\|\mathbb{P})^2 \equiv \sup_{\substack{h\in\mathcal{H}^d \\ \|h\|\leq 1}} \left| \int_{\mathcal{X}} \mathcal{S}_p^m[h] \mathrm{d}\mathbb{Q} \right|^2 = \int_{\mathcal{X}} \int_{\mathcal{X}} k^0(x, y) \mathrm{d}\mathbb{Q}(x) \mathrm{d}\mathbb{Q}(y) \tag{6.5}$$

$$k^0(x, y) \equiv \mathcal{S}_p^{m,2}\mathcal{S}_p^{m,1} K(x, y) = \frac{1}{p(y)p(x)} \nabla_y \cdot \nabla_x \cdot \big(p(x)m(x)K(x, y)m(y)^\top p(y)\big) . \tag{6.6}$$

As the proof shows, the Stein kernel k^0 is indeed a (scalar) kernel obtained from the feature map $\phi : \mathcal{X} \to \mathcal{H}^d$, $\phi(x) \equiv \mathcal{S}_p^{m,1}[K]|_x$. For $K = k\mathrm{Id}$, $m = h\mathrm{Id}$, DKSD is a KSD with scalar kernel $h(x)k(x, y)h(y)$, and if $h = 1$ our objective becomes the usual Langevin-based KSD of [83, 135, 226, 282] (see §A.7.3). The work of [228] discussed the potential of optimizing the KSD with (Euclidean) gradient descent but did not evaluate its merits. In the sections to follow, we will see the advantages conferred by introducing more flexible diffusion operators, matrix

kernels, and natural gradient optimisation.

Note it is also possible to view $K_m : (x, y) \mapsto m(x)K(x, y)m(y)^\top$ as a new matrix kernel, since $K_m(y, x)^\top = m(x)K(y, x)m(y)^\top = K_m(x, y)$, and for any $v_j \in \mathbb{R}^d, x_i \in \mathcal{X}$,

$$v_j \cdot K_m(x_j, x_l)v_l = v_j \cdot m(x_j)K(x_j, x_l)m(x_l)^\top v_l = \left(m(x_j)^\top v_j\right) \cdot K(x_j, x_l) \left(m(x_l)^\top v_l\right) \geq 0,$$

which shows K_m defines a vector RKHS (see §5.3). In order to use these for minimum SD estimation, we propose the following U-statistic approximation:

$$\widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n \| \mathbb{P}_\theta)^2 = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} k_\theta^0(X_i, X_j) = \frac{1}{n(n-1)} \sum_{i \neq j} k_\theta^0(X_i, X_j), \quad (6.7)$$

with associated estimators: $\hat{\theta}_n^{\mathrm{DKSD}} \in \mathrm{argmin}_{\theta \in \Theta} \widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n \| \mathbb{P}_\theta)^2$.

Now that our DKSD estimators are defined, an important remaining question is under which conditions can DKSD discriminate distinct probability measures. To answer, we will need several definitions. We say a matrix kernel K is in the Stein class of $\mathbb{Q}$ if $\int_{\mathcal{X}} \mathcal{S}_q^{m,1}[K] \mathrm{d}\mathbb{Q} = 0$, and that it is strictly integrally positive definite (IPD) if $\int_{\mathcal{X} \times \mathcal{X}} \mathrm{d}\mu^\top(x) K(x, y) \mathrm{d}\mu(y) > 0$ for any finite non-zero signed vector Borel measure μ. Note that when $K = k\mathrm{Id}$ we recover definition 3.4 [229] of Stein class of a scalar kernel with respect to $\mathbb{Q}$, as well as the definition in [326] of IPD scalar kernels. From $\mathcal{S}_p^m[f](x) = \langle \mathcal{S}_p^{m,1} K_x, f \rangle_{\mathcal{H}^d}$ we have that $f \in \mathcal{H}^d$ is in the Stein class (i.e., $\int_{\mathcal{X}} \mathcal{S}_q^m[f] \mathrm{d}\mathbb{Q} = 0$) when K is also in the class. Setting $s_p \equiv m^\top \nabla \log p \in \Gamma(\mathbb{R}^d)$:

Proposition 36 (DKSD as a Statistical Divergence). *Suppose K is IPD and in the Stein class of $\mathbb{Q}$, and $m(x)$ is invertible. If $s_p - s_q \in L^1(\mathbb{Q})$, then $\mathrm{DKSD}_{K,m}(\mathbb{Q} \| \mathbb{P})^2 = 0$ iff $\mathbb{Q} = \mathbb{P}$.*

See §6.11.2 for the proof. Note that this proposition generalises Proposition 3.3 from [226] to a significantly larger class of SD. For the matrix kernels introduced above, the proposition below shows that K is IPD when its associated scalar kernels are; a well-studied problem [325].

Proposition 37 (IPD Matrix Kernels). *(i) Let $K = diag(k^1, \ldots, k^d)$. Then K is IPD iff each kernel k^i is IPD. (ii) Let $K = Bk$ for B be symmetric positive definite. Then K is IPD iff k is IPD.*

6.4 Example 2: Diffusion Score Matching Estimators

A well-known family of estimators for sufficiently smooth densities are the SM estimators §6.1. We consider the approximation $\widehat{\mathrm{SM}}(\{X_i\}_{i=1}^n \| \mathbb{P}_\theta) \equiv \frac{1}{n} \sum_{i=1}^n \Delta \log p_\theta(X_i) + \frac{1}{2} \| \nabla \log p_\theta(X_i) \|_2^2$

based on an unbiased estimation for the minimiser of the SM discrepancy (6.2), and its estimators $\hat{\theta}_n^{\mathrm{SM}} \equiv \mathrm{argmin}_{\theta\in\Theta}\widehat{\mathrm{SM}}(\{X_i\}_{i=1}^n\|\mathbb{P}_\theta)$, for independent random vectors $X_i \sim \mathbb{Q}$. We will now introduce a generalisation that we call **diffusion score matching** (DSM) which is a SD constructed from the diffusion Stein operator (see §6.11.3):

Theorem 38 (**Diffusion Score Matching**). *Let $\mathcal{X} = \mathbb{R}^d$ and consider some diffusion Stein operator $\mathcal{S}_p^m$ for some function $m \in \Gamma(\mathbb{R}^{d\times d})$ and the Stein class $\mathcal{G} \equiv \{g = (g_1, \ldots, g_d) \in C^1(\mathcal{X}, \mathbb{R}^d) \cap L^2(\mathcal{X}; \mathbb{Q}) : \|g\|_{L^2(\mathcal{X};\mathbb{Q})} \leq 1\}$. If $p, q > 0$ are differentiable and $s_p - s_q \in L^2(\mathbb{Q})$, then we define the **diffusion score matching** divergence as the Stein discrepancy,*

$$\mathrm{DSM}_m(\mathbb{Q}\|\mathbb{P}) \equiv \sup_{f\in\mathcal{S}_p[\mathcal{G}]} \left|\int_{\mathcal{X}} f\mathrm{d}\mathbb{Q} - \int_{\mathcal{X}} f\mathrm{d}\mathbb{P}\right|^2 = \int_{\mathcal{X}} \left\|m^\top (\nabla \log q - \nabla \log p)\right\|_2^2 \mathrm{d}\mathbb{Q}. \tag{6.8}$$

This satisfies $\mathrm{DSM}_m(\mathbb{Q}\|\mathbb{P}) = 0$ iff $\mathbb{Q} = \mathbb{P}$ when $m(x)$ is invertible. Moreover, if p is twice-differentiable, and $qmm^\top \nabla \log p, \nabla \cdot (qmm^\top \nabla \log p) \in L^1(\mathbb{R}^d)$, then Stoke's theorem gives

$$\mathrm{DSM}_m(\mathbb{Q}\|\mathbb{P}) = \int_{\mathcal{X}} \left(\|m^\top \nabla_x \log p\|_2^2 + \|m^\top \nabla \log q\|_2^2 + 2\nabla \cdot \left(mm^\top \nabla \log p\right)\right) \mathrm{d}\mathbb{Q}. \tag{6.9}$$

Notably, DSM_m recovers SM when $m(x)m(x)^\top = I$, that is when m takes value in the orthogonal group, the group of linear maps that preserve the Euclidean inner product. We can also recover the (generalised) non-negative score matching estimator of [239] with the choice $m(x) \equiv \mathrm{diag}(h_1(x_1)^{1/2}, \ldots, h_d(x_d)^{1/2})$. As for the standard SM estimator, DSM is only defined for distributions with sufficiently smooth densities. However the θ-dependent part of $\mathrm{DSM}_m(\mathbb{Q}\|\mathbb{P}_\theta)$ in (6.9)[1]

$$\begin{aligned}
&\int_{\mathcal{X}} \left(\left\|m^\top \nabla_x \log p_\theta\right\|_2^2 + 2\nabla \cdot \left(mm^\top \nabla \log p_\theta\right)\right) \mathrm{d}\mathbb{Q} \\
&= \int_{\mathcal{X}} \left(\left\|m^\top \nabla_x \log p_\theta\right\|_2^2 + 2\left(\left\langle \nabla \cdot (mm^\top), \nabla \log p\right\rangle + \mathrm{tr}\left[mm^\top \nabla^2 \log p\right]\right)\right) \mathrm{d}\mathbb{Q},
\end{aligned}$$

or in components

$$\begin{aligned}
\int_{\mathcal{X}} \mathrm{d}\mathbb{Q}(x) \Big(& m(x)_{ij} m(x)_{lj} \partial_{x^l} \log p_\theta(x) \partial_{x^i} \log p_\theta(x) + 2\partial_{x^j}\partial_{x^k} \log p_\theta(x) m_{kl}(x) m_{jl}(x) \\
& + 2\partial_{x^k} \log p_\theta(x) \left(\partial_{x^j} m_{kl}(x) m_{jl}(x) + m_{kl}(x) \partial_{x^j} m_{jl}(x)\right) \Big),
\end{aligned}$$

does not depend on the density of $\mathbb{Q}$. Hence it can be estimated using an empirical mean,

[1] Here we use $\nabla \cdot \left(mm^\top \nabla \log p\right) = \left\langle \nabla \cdot (mm^\top), \nabla \log p\right\rangle + \mathrm{tr}\left[mm^\top \nabla^2 \log p\right]$

leading to the estimators $\hat{\theta}_n^{\mathrm{DSM}} \equiv \operatorname{argmin}_{\theta\in\Theta} \widehat{\mathrm{DSM}}_m(\{X_i\}_{i=1}^n \| \mathbb{P}_\theta)$ for

$$\widehat{\mathrm{DSM}}_m(\{X_i\}_{i=1}^n \| \mathbb{P}_\theta) \equiv \frac{1}{n}\sum_{i=1}^{n} \left(\|m^\top \nabla_x \log p_\theta\|_2^2 + 2\nabla \cdot \left(mm^\top \nabla \log p_\theta\right)\right)(X_i) \tag{6.10}$$

where $\{X_i\}_{i=1}^n$ is a sample from $\mathbb{Q}$. Note that this is only possible if m is independent of θ, in contrast to DKSD where m can depend on $\mathcal{X} \times \Theta$, thus leading to a more flexible class of estimators.

An interesting remark is that the DSM_m discrepancy may in fact be obtained as a limit of DKSD over a sequence of target-dependent kernels: see §6.11.3 for the complete result which corrects and significantly generalises previously established connections between the SM divergence and KSD (such as in Sec. 5 of [226]).

We conclude by commenting on the computational complexity. Evaluating the DKSD loss function requires $\mathcal{O}(n^2d^2)$ computation, due to the U-statistic and a matrix-matrix product. However, if $K = \mathrm{diag}(\lambda_1 k^1, \ldots, \lambda_d k^d)$ or $K = Bk$, and if m is a diagonal matrix, then we can by-pass expensive matrix products and the cost is $\mathcal{O}(n^2 d)$, making it comparable to that of KSD. Although we do not consider these in this paper, recent approximations to KSD could also be adapted to DKSD to reduce the computational cost to $\mathcal{O}(nd)$ [170, 178]. The DSM loss function has computational cost $\mathcal{O}(nd^2)$, which is comparable to the SM loss. From a computational viewpoint, DSM will hence be preferable to DKSD for large n, whilst DKSD will be preferable to DSM for large d.

6.5 Further Examples: Contrastive Divergence and Minimum Probability Flow

Before analysing DKSD and DSM estimators further, we show that the class of minimum SD estimators also includes other well-known estimators for unnormalised models. Let $(X_\theta^m)_{m\in\mathbb{N}}$ be an ergodic Markov chain with (unique) invariant probability measure $\mathbb{P}_\theta$, obtained for example using a Metropolis-Hastings algorithm §2.1. Let P_θ^m be the associated **transition semigroup**, i.e., $(P_\theta^m f)(x) \equiv \mathbb{E}[f(X_\theta^m)|X_\theta^0 = x]$ (see §4.1). Choosing the Stein operator $\mathcal{S}_{\mathbb{P}_\theta} \equiv I - P_\theta^m$, note that (5.8) holds (using (2.30-2.32) [294])

$$\int \mathcal{S}_{\mathbb{P}_\theta}(f)\mathrm{d}\mathbb{P}_\theta = \int f \mathrm{d}\mathbb{P}_\theta - \int P_\theta^m(f)\mathrm{d}\mathbb{P}_\theta = \int f\mathrm{d}\mathbb{P}_\theta - \int f\mathrm{d}\mathbb{P}_\theta = 0.$$

Using the Stein class $\{\log p_\theta + c : c \in \mathbb{R}\}$ leads to the following SD (see (6.3)):

$$\begin{aligned}\mathrm{SD}_{\mathcal{S}_{\mathbb{P}_\theta}}(\mathbb{Q}\|\mathbb{P}_\theta) &= \left|\int (\log p_\theta - P_\theta^m \log p_\theta)\,\mathrm{d}\mathbb{Q}\right| = \left|\int \log p_\theta \mathrm{d}\mathbb{Q} - \int \log p_\theta \mathrm{d}\mathbb{Q}_\theta^m\right| \\ &= \mathrm{KL}(\mathbb{Q}\|\mathbb{P}_\theta) - \mathrm{KL}(\mathbb{Q}_\theta^m\|\mathbb{P}_\theta) = \mathrm{CD}_m(\mathbb{Q}\|\mathbb{P}_\theta),\end{aligned}$$

where $\mathbb{Q}_\theta^m$ is the law of X_θ^m given that $X_\theta^0 \sim \mathbb{Q}$ and assuming that $\mathbb{Q} \ll \mathbb{P}_\theta$ and $\mathbb{Q}_\theta^m \ll \mathbb{P}_\theta$, which is the loss function associated with contrastive divergence (see (6.1)). Suppose now that $\mathcal{X}$ is a finite set. Given $\theta \in \Theta$ let P_θ be the transition matrix for a Markov process with unique invariant distribution $\mathbb{P}_\theta$. Suppose we observe data $\{x_i\}_{i=1}^n$ and let q be the corresponding empirical distribution. Choosing the Stein operator $\mathcal{S}_p = I - P_\theta$ and the Stein set $\mathcal{G} = \{f \in \Gamma(\mathbb{R}) : \|f\|_\infty \leq 1\}$. Note that, $g \in \arg\sup_{g\in G} |\mathbb{Q}(\mathcal{S}_p[g])|$ will satisfy $g(i) = \mathrm{sgn}(q^\top(I - P_\theta)_i)$, and the resulting Stein discrepancy is the minimum probability flow loss objective function [320]:

$$\mathrm{MPFL}(\mathbb{Q}\|\mathbb{P}) = \sum_y \left|((I - P_\theta)^\top q)_y\right| = \sum_{y \notin \{x_i\}_{i=1}^n} \left|\frac{1}{n} \sum_{x \in \{x_i\}_{i=1}^n} (I - P_\theta)_{xy}\right|. \tag{6.11}$$

6.6 Implementing Minimum SD Estimators: Stochastic Riemannian Gradient Descent

In order to implement the minimum SD estimators, we propose to use a stochastic gradient descent (SGD) algorithm associated to the information geometry induced by the SD on the parameter space §5.1. More precisely, consider a parametric family $\mathcal{P}_\Theta$ of probability measures on $\mathcal{X}$ with $\Theta \subset \mathbb{R}^m$. Indeed recall from §5.1.3 that a statistical divergence D on a parametrised statistical manifold has an associated information matrix field on Θ defined as the map $\theta \mapsto g(\theta)$, where $g(\theta)$ is the symmetric bilinear form $g(\theta)_{ij} = -\frac{1}{2}(\partial^2/\partial\alpha^i\partial\theta^j)D(\mathbb{P}_\alpha\|\mathbb{P}_\theta)|_{\alpha=\theta}$. When g is positive definite (i.e., D is a divergence) we can use it to perform (Riemannian) gradient descent on the parameter space Θ. We provide below the information matrices of DKSD and DSM (and hence extends results of [183]):

Proposition 39 (**Information Tensor DKSD**). *Assume the conditions of §36 hold. The information tensor associated to DKSD is positive semi-definite and has components*

$$g_{\mathrm{DKSD}}(\theta)_{ij} = \int_{\mathcal{X}}\int_{\mathcal{X}} (\nabla_x \partial_{\theta^j} \log p_\theta(x))^\top m_\theta(x) K(x,y) m_\theta^\top(y) \nabla_y \partial_{\theta^i} \log p_\theta(y) \mathrm{d}\mathbb{P}_\theta(x) \mathrm{d}\mathbb{P}_\theta(y). \tag{6.12}$$

Proposition 40 (**Information Tensor DSM**). *Assume the conditions of §38 hold. The information tensor defined by DSM is positive semi-definite and has components*

$$g_{\mathrm{DSM}}(\theta)_{ij} = \int_{\mathcal{X}} \left\langle m^\top \nabla \partial_{\theta^i} \log p_\theta, m^\top \nabla \partial_{\theta^j} \log p_\theta \right\rangle \mathrm{d}\mathbb{P}_\theta. \tag{6.13}$$

See §6.12 for the proofs. Following §5.2, given a sequence (η_t) of step sizes we will approximate the gradient flow with

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \eta_t \hat{g}_{\theta_t}(\{X_i^t\}_{i=1}^n)^{-1} \partial_{\theta_t} \widehat{\mathrm{SD}}(\{X_i^t\}_{i=1}^n \| \mathbb{P}_\theta). \tag{6.14}$$

Minimum SD estimators hold additional appeal for exponential family models, since their densities have the form $p_\theta(x) \propto \exp\left(\langle \theta, T(x) \rangle_{\mathbb{R}^m}\right) \exp(b(x))$ for natural parameters $\theta \in \mathbb{R}^m$, sufficient statistics $T \in \Gamma(\mathbb{R}^m)$, and base measure $\exp(b(x))$. For these models, the U-statistic approximations of DKSD and DSM are convex quadratics with closed form solutions whenever K and m are independent of θ. Moreover, since the absolute value of an affine function is convex, and the supremum of convex functions is convex, any SD with a diffusion Stein operator is convex in θ, provided m and the Stein class $\mathcal{G}$ are independent of θ.

6.7 Theoretical Properties for Minimum Stein Discrepancy Estimators

We now show that the DKSD and DSM estimators have many desirable properties such as consistency, asymptotic normality and bias-robustness. These results do not only provide us with reassuring theoretical guarantees on the performance of our algorithms, but can also be a practical tool for choosing a Stein operator and Stein class given an inference problem of interest.

We begin by establishing strong consistency and for DKSD; i.e. almost sure convergence: $\hat{\theta}_n^{\mathrm{DKSD}} \xrightarrow{a.s.} \theta_*^{\mathrm{DKSD}} \equiv \operatorname{argmin}_{\theta \in \Theta} \mathrm{DKSD}_{K,m}(\mathbb{Q} \| \mathbb{P}_\theta)^2$. This will be followed by a proof of asymptotic normality. We will assume we are in the specified setting, so that $\mathbb{Q} = \mathbb{P}_{\theta_*^{\mathrm{DKSD}}} \in \mathcal{P}_\Theta$. In the misspecified setting, we will need to also assume the existence of a unique minimiser.

Theorem 41 (**Strong Consistency of DKSD**). *Let $\mathcal{X} = \mathbb{R}^d$, $\Theta \subset \mathbb{R}^m$. Suppose that K is bounded with bounded derivatives up to order 2, that $k^0(x, y)$ is continuously-differentiable on an $\mathbb{R}^m$-open neighbourhood of Θ, and that for any compact subset $C \subset \Theta$ there exist functions f_1, f_2, g_1, g_2 such that for $\mathbb{Q}$-a.e. $x \in \mathcal{X}$,*

1. $\left\|m^\top(x)\nabla \log p_\theta(x)\right\| \leq f_1(x)$, *where* $f_1 \in L^1(\mathbb{Q})$ *and continuous,*

2. $\left\|\nabla_\theta \left(m(x)^\top \nabla \log p_\theta(x)\right)\right\| \leq g_1(x)$, *where* $g_1 \in L^1(\mathbb{Q})$ *is continuous,*

3. $\|m(x)\| + \|\nabla_x m(x)\| \leq f_2(x)$ *where* $f_2 \in L^1(\mathbb{Q})$ *and continuous,*

4. $\|\nabla_\theta m(x)\| + \|\nabla_\theta \nabla_x m(x)\| \leq g_2(x)$ *where* $g_2 \in L^1(\mathbb{Q})$ *is continuous.*

Assume further that $\theta \mapsto \mathbb{P}_\theta$ *is injective. Then we have a unique minimiser* θ_*^{DKSD}*, and if either* Θ *is compact, or* $\theta_*^{\mathrm{DKSD}} \in int(\Theta)$ *and* Θ *and* $\theta \mapsto \widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n \| \mathbb{P}_\theta)^2$ *are convex, then* $\hat{\theta}_n^{\mathrm{DKSD}}$ *is strongly consistent.*

Theorem 42 (**Central Limit Theorem for DKSD**). *Let* $\mathcal{X}$ *and* Θ *be open subsets of* $\mathbb{R}^d$ *and* $\mathbb{R}^m$ *respectively. Let* K *be a bounded kernel with bounded derivatives up to order* 2 *and suppose that* $\hat{\theta}_n^{\mathrm{DKSD}} \xrightarrow{p} \theta_*^{\mathrm{DKSD}}$ *and that there exists a compact neighbourhood* $\mathcal{N} \subset \Theta$ *of* θ_*^{DKSD} *such that* $\theta \to \widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_\theta)^2$ *is twice continuously differentiable for* $\theta \in \mathcal{N}$ *and, for* $\mathbb{Q}$*-a.e.* $x \in \mathcal{X}$*,*

1. $\|m^\top(x)\nabla \log p_\theta(x)\| + \|\nabla_\theta \left(m(x)^\top \nabla \log p_\theta(x)\right)\| \leq f_1(x)$,

2. $\|m(x)\| + \|\nabla_x m(x)\| + \|\nabla_\theta m(x)\| + \|\nabla_\theta \nabla_x m(x)\| \leq f_2(x)$,

3. $\|\nabla_\theta \nabla_\theta \left(m(x)^\top \nabla \log p_\theta(x)\right)\| + \|\nabla_\theta \nabla_\theta \nabla_\theta \left(m(x)^\top \nabla \log p_\theta(x)\right)\| \leq g_1(x)$,

4. $\|\nabla_\theta \nabla_\theta m(x)\| + \|\nabla_\theta \nabla_\theta \nabla_x m(x)\| + \|\nabla_\theta \nabla_\theta \nabla_\theta m(x)\| + \|\nabla_\theta \nabla_\theta \nabla_\theta \nabla_x m(x)\| \leq g_2(x)$,

where $f_1, f_2 \in L^2(\mathbb{Q})$,$g_1, g_2 \in L^1(\mathbb{Q})$ *are continuous. Suppose also that the information tensor* g *is invertible at* θ_*^{DKSD}*. Then*

$$\sqrt{n}\left(\hat{\theta}_n^{\mathrm{DKSD}} - \theta_*^{\mathrm{DKSD}}\right) \xrightarrow{d} \mathcal{N}\left(0, g_{\mathrm{DKSD}}^{-1}(\theta_*^{\mathrm{DKSD}})\Sigma_{\mathrm{DKSD}} g_{\mathrm{DKSD}}^{-1}(\theta_*^{\mathrm{DKSD}})\right), \tag{6.15}$$

where $\Sigma_{\mathrm{DKSD}} = \int_{\mathcal{X}} \left(\int_{\mathcal{X}} \nabla_\theta k^0_{\theta_*^{\mathrm{DKSD}}}(x,y)\mathrm{d}\mathbb{Q}(y)\right) \otimes \left(\int_{\mathcal{X}} \nabla_\theta k^0_{\theta_*^{\mathrm{DKSD}}}(x,z)\mathrm{d}\mathbb{Q}(z)\right) \mathrm{d}\mathbb{Q}(x)$.

See §6.13 for proofs. For both results, the assumptions on the kernel are satisfied by most kernels common in the literature, such as Gaussian, inverse-multiquadric (IMQ) and any Matérn kernels with smoothness greater than 2. Similarly, the assumptions on the model are very weak given that the diffusion tensor m can be adapted to guarantee consistency and asymptotic normality.

We now prove analogous results for DSM. This time we show weak consistency, i.e. convergence in probability: $\hat{\theta}_n^{\mathrm{DSM}} \xrightarrow{p} \theta_*^{\mathrm{DSM}} \equiv \mathrm{argmin}_{\theta\in\Theta} \mathrm{DSM}_m(\mathbb{Q}\|\mathbb{P}_\theta)$. This will be a sufficient form of convergence for asymptotic normality (see (6.23a) for definition of F_θ).

Theorem 43 (**Weak Consistency of DSM**). *Let $\mathcal{X}$ be an open subset of $\mathbb{R}^d$, and $\Theta \subset \mathbb{R}^m$. Suppose $\log p_\theta(\cdot) \in C^2(\mathcal{X})$ and $m \in C^1(\mathcal{X})$, and $\|\nabla_x \log p_\theta(x)\| \leq f_1(x)$ for $\mathbb{Q}$-a.e. x. Suppose also that $\|\nabla_x \nabla_x \log p_\theta(x)| \leq f_2(x)$ on any compact set $C \subset \Theta$ for $\mathbb{Q}$-a.e. x, where $\|m^\top\| f_1 \in L^2(\mathbb{Q})$, $\|\nabla \cdot (mm^\top)\| f_1 \in L^1(\mathbb{Q})$, $\|mm^\top\|_\infty f_2 \in L^1(\mathbb{Q})$. If either Θ is compact, or Θ and $\theta \mapsto F_\theta$ are convex and $\theta_*^{\mathrm{DSM}} \in int(\Theta)$, then $\hat{\theta}_n^{\mathrm{DSM}}$ is weakly consistent for θ_*^{DSM}.*

When weak consistency holds and Θ is an open subset of Euclidean space, we can derive the following CLT

Theorem 44 (**Central Limit Theorem for DSM**). *Let $\mathcal{X}, \Theta$ be open subsets of $\mathbb{R}^d$ and $\mathbb{R}^m$ respectively. Suppose $\hat{\theta}_n^{\mathrm{DSM}} \xrightarrow{p} \theta_*^{\mathrm{DSM}}$, $\theta \mapsto \log p_\theta(x)$ is twice continuously differentiable on a closed ball $\bar{B}(\epsilon, \theta_*^{\mathrm{DSM}}) \subset \Theta$, and that for $\mathbb{Q}$-a.e. $x \in \mathcal{X}$,*

- ***(i)*** *$\|m(x)m^\top(x)\| + \|\nabla_x \cdot (m(x)m^\top(x))\| \leq f_1(x)$, and $\|\nabla_x \log p_\theta(x)\| + \|\nabla_\theta \nabla_x \log p_\theta(x)\| + \|\nabla_\theta \nabla_x \nabla_x \log p_\theta(x)\| \leq f_2(x)$, with $f_1 f_2, f_1 f_2^2 \in L^2(\mathbb{Q})$*
- ***(ii)*** *for $\theta \in \bar{B}(\epsilon, \theta^*)$, $\|\nabla_\theta \nabla_x \log p_\theta\|^2 + \|\nabla_x \log p_\theta\| \|\nabla_\theta \nabla_\theta \nabla_x \log p_\theta\| + \|\nabla_\theta \nabla_\theta \nabla_x \log p_\theta\| + \|\nabla_\theta \nabla_\theta \nabla_x \nabla_x \log p_\theta\| \leq g_1(x)$, and $f_1 g_1 \in L^1(\mathbb{Q})$.*

Then, if the information tensor is invertible at θ_*^{DSM}, we have

$$\sqrt{n}\left(\hat{\theta}_n^{\mathrm{DSM}} - \theta_*^{\mathrm{DSM}}\right) \xrightarrow{d} \mathcal{N}\left(0, g_{\mathrm{DSM}}^{-1}\left(\theta_*^{\mathrm{DSM}}\right) \Sigma_{\mathrm{DSM}} g_{\mathrm{DSM}}^{-1}\left(\theta_*^{\mathrm{DSM}}\right)\right). \tag{6.16}$$

where $\Sigma_{\mathrm{DSM}} = \int_{\mathcal{X}} \nabla_\theta F_{\theta_*^{\mathrm{DSM}}}(x) \otimes \nabla_\theta F_{\theta_*^{\mathrm{DSM}}}(x) \mathrm{d}\mathbb{Q}(x)$.

All of the proofs can be found in §6.13.2. An important special case covered by our theory is that of natural exponential families, which admit densities of the form $\log p_\theta(x) \propto \langle \theta, T(x) \rangle_{\mathbb{R}^m} + b(x)$. If K is IPD with bounded derivative up to order 2, ∇T has linearly independent rows, m is invertible, and $\|\nabla T m\|, \|\nabla_x b\| \|m\|, \|\nabla_x m\| + \|m\| \in L^2(\mathbb{Q})$, then the sequence of minimum DKSD and DSM estimators are strongly consistent and asymptotically normal (see §6.13.3).

Before concluding this section, we turn to a concept of importance to practical inference: robustness when subjected to corrupted data [169]. We quantify the robustness of DKSD and DSM estimators in terms of their influence function, which can be interpreted as measuring the impact of an infinitesimal perturbation of a distribution $\mathbb{P}$ by a Dirac located at a point $z \in \mathcal{X}$ on the estimator. If $\theta_{\mathbb{Q}}$ denotes the unique minimum SD estimator for $\mathbb{Q}$, then the **influence functions** is given by

$$\mathrm{IF}(z, \mathbb{Q}) \equiv \partial_t \theta_{\mathbb{Q}_t}|_{t=0}$$

if it exists, where $\mathbb{Q}_t = (1-t)\mathbb{Q} + t\delta_z$, for $t \in [0, 1]$. An estimator is said to be **bias robust** if $z \mapsto \mathrm{IF}(z, \mathbb{Q})$ is bounded.

Proposition 45 (Robustness of DKSD estimators). *Suppose that the map $\theta \to \mathbb{P}_\theta$ over Θ is injective, then* $\mathrm{IF}(z, \mathbb{P}_\theta) = g_{\mathrm{DKSD}}(\theta)^{-1} \int_X \nabla_\theta k^0(z, y) \mathrm{d}\mathbb{P}_\theta(y)$. *Moreover, suppose that $y \mapsto F(x, y)$ is $\mathbb{Q}$-integrable for any x, where $F(x, y) = \|K(x, y)s_p(y)\|$, $\|K(x, y)\nabla_\theta s_p(y)\|$, $\|\nabla_x K(x, y)s_p(y)\|$, $\|\nabla_x K(x, y)\nabla_\theta s_p(y)\|$, $\|\nabla_y \nabla_x (K(x, y)m(y))\|$,$\|\nabla_y \nabla_x (K(x, y)\nabla_\theta m(y))\|$. Then if $x \mapsto (\|s_p(x)\| + \|\nabla_\theta s_p(x)\|) \int F(x, y)\mathbb{Q}(dy)|_{\theta_*^{\mathrm{DKSD}}}$ is bounded, the* DKSD *estimators are bias robust:* $\sup_{z\in\mathcal{X}} \|\mathrm{IF}(z, \mathbb{Q})\| < \infty$.

The analogous results for DSM estimators can be found in §6.14. Consider a Gaussian location model, i.e. $p_\theta \propto \exp(-\|x - \theta\|_2^2)$, for $\theta \in \mathbb{R}^d$. The Gaussian kernel satisfies the assumptions of proposition 45 so that $\sup_z \|\mathrm{IF}(z, \mathbb{Q})\| < \infty$, even when $m = I$. Indeed $\|\mathrm{IF}(z, \mathbb{P}_\theta)\| \leq C(\theta)e^{-\|z-\theta\|^2/4}\|z - \theta\|$, where $z \mapsto e^{-\|z-\theta\|^2/4}\|z - \theta\|$ is uniformly bounded over θ. In contrast, the SM estimator has an influence function of the form $\mathrm{IF}(z, \mathbb{Q}) = z - \int_{\mathcal{X}} x \mathrm{d}\mathbb{Q}(x)$, which is unbounded with respect to z, and is thus not robust. This clearly demonstrates the importance of carefully selecting a Stein class for use in minimum SD estimators. An alternative way of inducing robustness is to introduce a spatially decaying diffusion matrix in DSM. To this end, consider the minimum DSM estimator with scalar diffusion coefficient m. Then $\theta_{\mathrm{DSM}} = (\int_{\mathcal{X}} m^2(x)\mathrm{d}\mathbb{Q}(x))^{-1} \left(\int_{\mathcal{X}} m^2(x)x\mathrm{d}\mathbb{Q}(x) + \int_{\mathcal{X}} \nabla m^2(x)\mathrm{d}\mathbb{Q}(x)\right)$. A straightforward calculation yields that the associated influence function will be bounded if both $m(x)$ and $\|\nabla m(x)\|$ decay as $\|x\| \to \infty$. This clearly demonstrates another significant advantage provided by the flexibility of our family of diffusion SD, where the Stein operator also plays an important role.

6.8 Numerical Experiments

In this section, we include the numerical experiments of [27] which explore several examples which demonstrate worrying breakpoints for SM, and highlight how these can be straightforwardly handled using KSD, DKSD and DSM. The calculations for the experiments were done by myself, and the experiments were done by Doctor François-Xavier Briol for our paper [27].

6.8.1 Rough densities: the symmetric Bessel distributions

A major drawback of SM is the smoothness requirement on the target density. However, this can be remedied by choosing alternative Stein classes, as will be demonstrated in the case of the symmetric Bessel distributions. Let $K_{s-d/2}$ denote the modified Bessel function of the second kind with parameter $s - d/2$. This distribution generalises the Laplace distribution [198] and has log-density: $\log p_\theta(x) \propto (\|x - \theta_1\|_2/\theta_2)^{(s-d/2)} K_{s-d/2}(\|x - \theta_1\|_2/\theta_2)$ where $\theta_1 \in \mathbb{R}^d$ is a

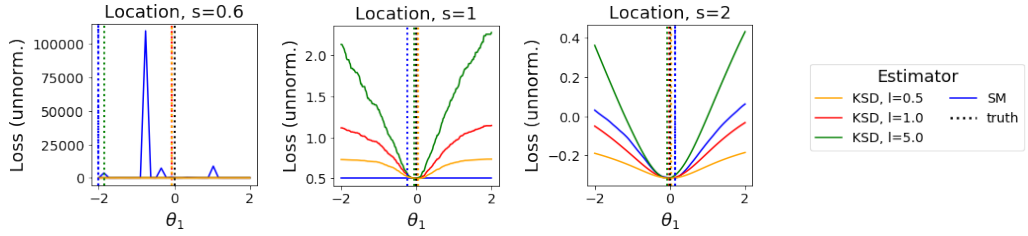

Figure 6.1: *Minimum SD Estimators for the Symmetric Bessel Distribution.* We consider the case where $\theta_1^* = 0$ and $\theta_2^* = 1$ and $n = 500$ for a range of smoothness parameter values s in $d = 1$.

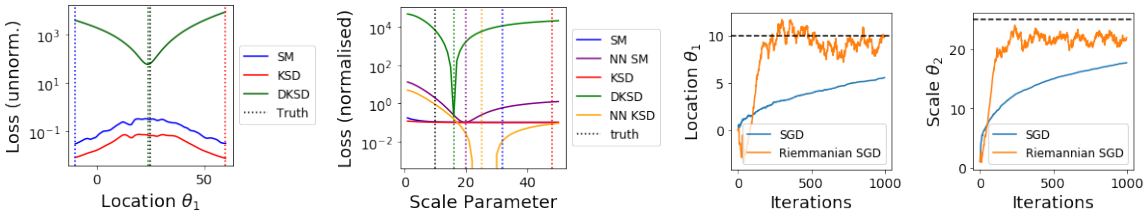

Figure 6.2: *Minimum SD Estimators for Non-standardised Student-t Distributions.* We consider a student-t problem with $\nu = 5, \theta_1^* = 25, \theta_2^* = 10$ and $n = 300$.

location parameter and $\theta_2 > 0$ a scale parameter. The parameter $s \geq d/2$ encodes smoothness.

We compared SM with KSD based on a Gaussian kernel and a range of lengthscale values in 6.1. These results are based on $n = 500$ IID realisations in $d = 1$. The case $s = 1$ corresponds to a Laplace distribution, and we notice that both SM and KSD are able to obtain a reasonable estimate of the location. For rougher values, for example $s = 0.6$, we notice that KSD outperforms SM for certain choices of lengthscales, whereas for $s = 2$, SM and KSD are both able to recover the parameter. Analogous results for scale can be found in 6.8.5, and 6.8.6 illustrates the trade-off between efficiency and robustness on this problem.

6.8.2 Heavy-tailed distributions: the non-standardised student-t distributions

A second drawback of standard SM is that it is inefficient for heavy-tailed distributions. To demonstrate this, we focus on non-standardised student-t distributions: $p_\theta(x) \propto (1/\theta_2)(1 + (1/\nu)\|x - \theta_1\|_2^2/\theta_2^2)^{-(\nu+1)/2}$ where $\theta_1 \in \mathbb{R}$ is a location parameter and $\theta_2 > 0$ a scale parameter. The parameter ν determines the degrees of freedom: when $\nu = 1$, we have a Cauchy distribution, whereas $\nu = \infty$ gives the Gaussian distribution. For small values of ν, the student-t distribution is heavy-tailed.

We illustrate SM and KSD for $\nu = 5$ in Fig. 6.2, where we take an IMQ kernel $k(x, y; c, \beta) = (c^2 + \|x - y\|_2^2)^\beta$ with $c = 1.$ and $\beta = -0.5$. This choice of ν guarantees the first two moments

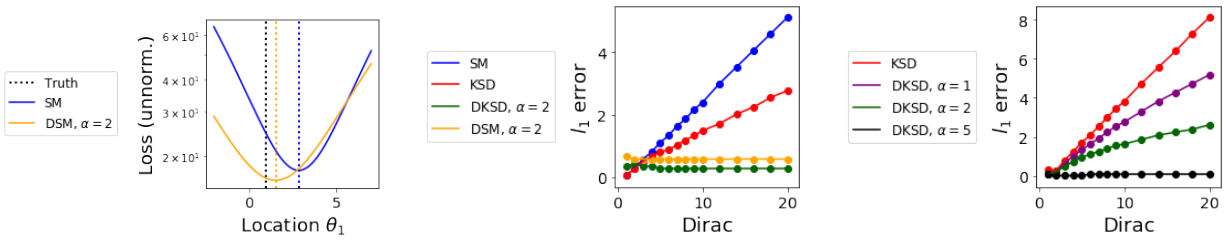

Figure 6.3: *Minimum SD Estimators for Generalised Gamma Distributions under Corruption.* We consider the case where $\theta_1^* = 0$ and $\theta_2^* = 2$ (left and middle) or $\theta_2^* = 5$ (right). Here $n = 300$.

exist, but the distribution is still heavy-tailed. In the left plot, both SM and KSD struggle to recover θ_1^* when $n = 300$, and the loss functions are far from convex. However, DKSD with $m_\theta(x) = 1 + \|x - \theta_1\|^2/\theta_2^2$ can estimate θ_1 very accurately. In the middle left plot, we instead estimate θ_2 with SM, KSD and their correponding non-negative version (NNSM & NNKSD, $m(x) = x$), which are particularly well suited for scale parameters. NNSM and NNKSD provide improvements on SM and KSD, but DKSD with $m_\theta(x) = ((x - \theta_1)/\theta_2)(1 + (1/\nu)\|x - \theta_1\|_2^2/\theta_2^2)$ provides significant further gains. On the right-hand side, we also consider the advantage of the Riemannian SGD algorithm over SGD by illustrating them on the KSD loss function with $n = 1000$. Both algorithms use constant stepsizes and minibatches of size 50. As demonstrated, Riemmannian SGD converges within a few dozen iterations, whereas SGD hasn't converged after 1000 iterations. Additional experiments on the robustness of these estimators is also available in 6.8.6.

6.8.3 Robust estimators for light-tailed distributions: the generalised Gamma distributions

Our final example demonstrates a third failure mode for SM: its lack of robustness for light-tailed distributions. We consider generalised gamma location models with likelihoods $p_\theta(x) \propto \exp(-(x-\theta_1)^{\theta_2})$ where θ_1 is a location parameter and θ_2 determines how fast the tails decay. The larger θ_2, the lighter the tails will be and vice-versa. We set $n = 300$ and corrupt 80 points by setting them to the value $x = 8$. A robust estimator should obtain a good approximation of θ^* even under this corruption. The left plot in 6.3 considers a Gaussian model (i.e. $\theta_2^* = 2$); we see that SM is not robust for this very simple model whereas DSM with $m(x) = 1/(1+\|x\|^\alpha), \alpha = 2$ is robust. The middle plot shows that DKSD with this same m is also robust, and confirms the analytical results of the previous section. Finally, the right plot considers the case $\theta_2^* = 5$ and we see that α can be chosen as a function of θ_2 to guarantee robustness. In general, taking $\alpha \geq \theta_2^* - 1$ will guarantee a bounded influence function. Such a choice allows us to obtain robust estimators even for models with very light tails.

6.8.4 Efficient estimators for a simple unnormalised model

Finally we consider a simple intractable model from [231]: $p_\theta(x) \propto \exp(\eta(\theta)^\top \psi(x))$ where $\psi(x) = (\sum_{i=1}^d x_i^2, \sum_{i=3}^d x_1 x_i, \tanh(x))^\top$ and tanh is applied elementwise to x and $\eta(\theta) = (-0.5, 0.2, 0.6, 0, 0, 0, \theta, 0)$. This model is intractable since we cannot easily compute its normalisation constant due to the difficulty of integrating the unnormalised part of the model. Our results based on $n = 200$ samples show that DKSD with $m(x) = \text{diag}(1/(1+x))$ is able to recover $\theta^* = -1$, whereas both SM and KSD provide less accurate estimates of the parameter. This illustrates yet again that a judicious choice of diffusion matrix can significantly improve the efficiency of our estimators.

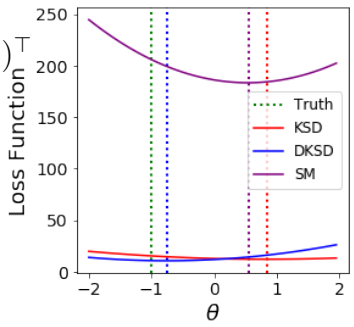

Figure 6.4: *Estimators for a Simple Intractable Model*

6.8.5 Efficiency of Minimum SD Estimators for Scale Parameters of Symmetric Bessel distributions

In this section, we extend the results from the main text and compares SM with KSD based on a Gaussian kernel and a range of lengthscale values for the scale parameter of the symmetric Bessel distribution. The results, given in 6.1, are also based on $n = 500$ IID realisations in $d = 1$. Similar results to those for the location parameter are obtained: KSD can deal with rougher densities, as illustrated when $s = 0.6$.

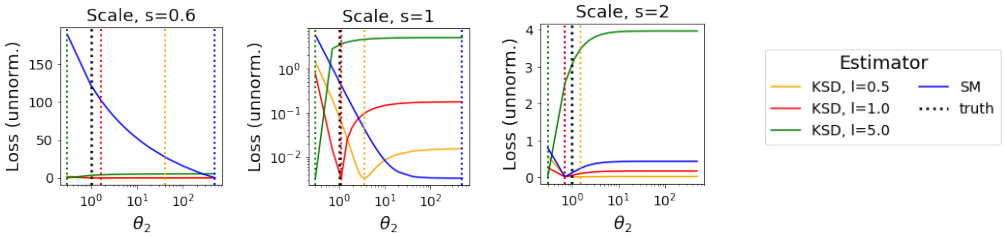

Figure 6.5: *Minimum SD Estimators for the Scale of a Symmetric Bessel Distribution.* We consider the case where $\theta_1^* = 0$ and $\theta_2^* = 1$ and $n = 500$ for a range of smoothness parameter values s in $d = 1$.

6.8.6 Bias Robustness of Minimum SD Estimators for the Symmetric Bessel and Non-standardised Student-t Distributions

In this section, we explore the robustness of minimum SD estimators for the two other examples in the main paper: the symmetric Bessel distribution ($\nu = 1000$) and the non-standardised student-t distribution. We once again select a diffusion matrix of the form $m(x) = 1/(1+\|x\|^{\alpha})$, and fix $\alpha = 1$ in both cases. This choice is refered to as "robust DKSD". On the other hand, we call "efficient DKSD" the DKSDs with choices of m as highlighted in the main text (and which were chosen to improve efficiency in both cases). The results are provided in 6.6. In each case, we used $n = 500$ data points, 80 of which were corrupted by a Dirac at some value of given on the x-axis. Both in the student-t and symmetric Bessel distribution, we notice that the "efficient DKSD" has an l_1 error which grows with the value of the Dirac, whereas the "robust DKSD" is bounded as a function of this Dirac.

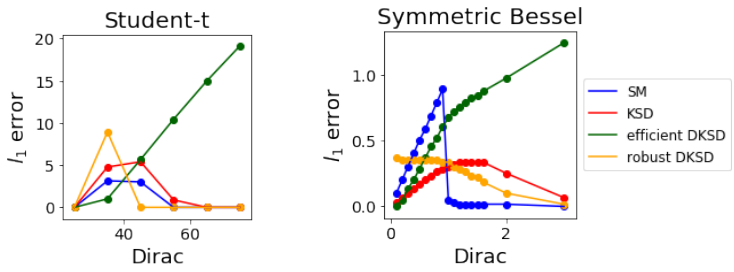

Figure 6.6: *The Robustness of Minimum SD Estimators for the Symmetric Bessel and Student-t Distributions. Left:* Student-t distribution. *Right:* Symmetric Bessel distribution.

6.9 Geometry of Stein Operators

Finally, let us go back to assume we have a smooth positive target distribution P to discuss Stein operators and the associated discrepancies from a geometric perspective. As mentioned earlier, Stein operators are used in statistics to construct functions with zero expectation with respect to some smooth positive target measure P. In the following we shall consider Stein operators acting on tensor fields, that is we want the Stein class to be a set of sections (of appropriate bundles). To understand what is the foundation of Stein operators, we begin by considering the integral with respect to P as a linear operator on test functions (i.e., as a generalised function)

$$P[\cdot] : C_c^{\infty}(\mathcal{X}) \to \mathbb{R}, \qquad Pf \equiv \int fP,$$

and are interested in obtaining an operator that generates test functions in the kernel of $P[\cdot]$. We have the following theorem based on the fact that when $\mathcal{X}$ is connected and oriented, there is an isomorphism between the top compactly supported top de Rham cohomology group $H_c^n(\mathcal{X})$ and $\mathbb{R}$ given by integration, $\omega \mapsto \int_{\mathcal{X}} \omega$, so in particular any compactly supported smooth measure whose integral vanishes is exact [209, Sec. 17]. As usual, while this theorem does not depend on the particular target measure P, the result of statistical interest is simply its transformation using the P-musical isomorphisms.

Theorem 46. *If $\mathcal{X}$ is connected and oriented, then the linear map $P[\cdot] : C_c^\infty(\mathcal{X}) \to \mathbb{R}$ descends to an isomorphism of vector spaces*

$$P[\cdot] : C_c^\infty(\mathcal{X})/Im(div_P|_c) \to \mathbb{R},$$

where $div_P|_c : \mathfrak{X}_c(\mathcal{X}) \to C_c^\infty(\mathcal{X})$ is the divergence operator restricted to compactly supported vector fields.

Proof. Indeed, the operator $P[\cdot]$ is the composition of two isomorphisms, the cohomology integral, $\int : H_c^n(\mathcal{X}) \to \mathbb{R}$, and the P-musical isomorphism restricted to the compactly supported 0^{th}-homology of P, $P^\flat : C_c^\infty(\mathcal{X})/\mathrm{Im}(\mathrm{div}_P|_c) \to H_c^n(\mathcal{X})$. ■

Hence, if $h, f \in C_c^\infty(\mathcal{X})$ are test functions, we have

$$\int fP = \int hP \quad \iff \quad f = h + \mathrm{div}_P(X), \text{ for some } X \in \mathfrak{X}_c(\mathcal{X}).$$

Thus, denoting by $[\cdot]_P : C_c^\infty(\mathcal{X}) \to C_c^\infty(\mathcal{X})/\mathrm{Im}(\mathrm{div}_P|_c)$ the projection onto the equivalence class of the homology of P, we have

$$(P[\cdot])^{-1}(0) = [0]_P = \{\mathrm{div}_P(X) : X \in \mathfrak{X}_c(\mathcal{X})\},$$

that is the test functions that integrate to 0 are precisely those that can be written as the divergence of compactly supported vector fields. In particular:

Corollary 47. *If $\mathcal{X}$ is compact, connected, and oriented, then viewing the integral operator with respect to P as an operator on smooth functions, we have*

$$\int fP = 0 \quad \iff \quad f = div_P(X) \text{ for some } X \in \mathfrak{X}(\mathcal{X}).$$

Thus, on compact manifolds the situation is clear, there is a canonical Stein operator, div_P, and a canonical Stein class, the space of vector fields $\mathfrak{X}(\mathcal{X})$. Interestingly, we see that this result

is a direct consequence of the properties of the top compactly supported de Rham cohomology group. More generally, the Stein operator and Stein class is provided by Stokes' theorem[2]

Theorem 48. *Let $\mathcal{X}$ be a connected and oriented manifold. Then, viewing integration as an operator on smooth P-integrable functions*

$$\int fP = 0 \quad \Longleftrightarrow \quad f = div_P(P^\sharp(\alpha)) \text{ where } \int \mathrm{d}\alpha = 0.$$

Again, it is interesting to observe that Stokes' theorem

$$\int \mathrm{d}\alpha = 0,$$

is a canonical statement about the smooth manifold $\mathcal{X}$, and does not require imposing any additional structure. As the above result shows, the vanishing property of Stein operators, which is sometimes the only one we need in applications, is simply a rephrasing of Stokes' theorem via the P-musical isomorphisms.

Hence, in general the Stein operator is the P-rotationnel restricted to vector fields in the **canonical Stein class**

$$\mathfrak{X}_P \equiv P^\sharp\left(\{\alpha \in \Omega^{n-1}(\mathcal{X}) : \int \mathrm{d}\alpha = 0\}\right).$$

We note that the canonical Stein class contains the P-preserving dynamics (since these are the kernel of div_P), and the set $\mathfrak{X}_{c,0}$ of vector fields X for which $i_X P$ is compactly supported and vanish on the boundary $\partial\mathcal{X}$ if the latter is non-empty.

We will thus call

$$\mathcal{T}_P \equiv \mathrm{div}_P|_{\mathfrak{X}_P} : \mathfrak{X}_P \to C^\infty(\mathcal{X})$$

the **canonical Stein operator**; by construction it satisfies the desired property

$$\mathbb{E}_P[\mathcal{T}_P(X)] \equiv \int \mathcal{T}_P(X)P = 0, \quad \forall X \in \mathfrak{X}_P.$$

As mentioned above, the canonical Stein operator is canonical in the sense that it does not require any additional structure on the sample space: it is fully determined by the target measure P. Moreover the canonical Stein operator only depends on P up to normalisation, since $\mathrm{curl}_{aP} = \mathrm{curl}_P$ for any $a \in \mathbb{R}/\{0\}$, an important requirement in many areas of statistics.

If we fix a bracket $\mathcal{B}$ on $\mathcal{X}$, the canonical Stein operator on vector fields induces a second order

[2]See for example [299, 72, 347] for discussions on Stokes' theorem on non-compact manifolds.

operator on functions, the $\mathcal{B}$-Stein operator on the Stein class $C_P^\infty \equiv \{f \in C^\infty(\mathcal{X}) : X_f \in \mathfrak{X}_P\}$,

$$\mathcal{T}_P^{\mathcal{B}} : C_P^\infty \to C^\infty(\mathcal{X}) \qquad \mathcal{T}_P^{\mathcal{B}} f \equiv \mathrm{div}_P(X_f) \equiv \mathrm{div}_{e^{-H}\mu_{\mathcal{X}}}(X_f) = \mathrm{div}_{\mu_{\mathcal{X}}}(X_f) - \{f, H\}.$$

- When $\mathcal{B} \equiv \mathcal{A}$ is antisymmetric, the Bracket-Stein operator is simply a first-order differential operator, the curl vector field of the target

$$\mathcal{T}_P^{\mathcal{A}} = \phi_P^{\mathcal{A}}.$$

- When $\mathcal{X}$ is Riemannian, $\mathcal{B} \equiv G^{-1}$, and $P = p_\infty \mathrm{vol} = e^{-H}\mathrm{vol}$, then

$$\mathcal{T}_P^G(f) = \nabla \cdot X_f - \langle \nabla f, \nabla H \rangle = \Delta f - \langle \nabla f, \nabla H \rangle = \Delta f + \langle \nabla f, \nabla \log p_\infty \rangle ,$$

  and the G^{-1}-Stein operator becomes the Riemannian Stein operator [29, 340, 154]. Hence the Riemannian Stein operator is the restriction of the canonical Stein operator to gradient fields. We proposed the Riemannian Stein operator in [29], where we analysed a method for numerical integration of smooth functions based on interpolation with the induced Riemannian Stein kernel. In particular, the resulting estimator for the integral was shown to be consistent and rate-optimal under appropriate assumptions, which not only extended but also improved previous results on Euclidean space [282, 281] (the proofs were done by myself, except the MCMC ones which were done by both Dr. Oates and I).
- In general, decomposing the bracket into its symmetric and antisymmetric part, $\mathcal{B} \equiv \mathcal{S} + \mathcal{A}$, we obtain the following useful decomposition of the $\mathcal{B}$-Stein operator

$$\mathcal{T}_P^{\mathcal{S}+\mathcal{A}}(f) = \mathrm{div}_P(X_f^{\mathcal{S}} + X_f^{\mathcal{A}}) = \phi_P^{\mathcal{S}}(f) + \phi_P^{\mathcal{A}}(f). \tag{6.17}$$

  When $\mathcal{S}$ is the canonical symmetric bracket induced by a finite set of vector fields we recover the generator of P-reserving diffusions (4.35).
- In fact the brackets allows us to define a more general notion of Stein operator acting on 1-forms in the Stein class $\Omega_P^1 \equiv \{\alpha \in \Omega^1(\mathcal{X}) : \mathcal{B}^\sharp(\alpha) \in \mathfrak{X}_P\}$, as the pullback under $\mathcal{B}^\sharp$ of the canonical Stein operator

$$\mathcal{T}_P^{\mathcal{B}} \equiv (\mathcal{B}^\sharp)^* \mathcal{T}_P : \Omega_P^1 \to C^\infty(\mathcal{X}), \qquad \mathcal{T}_P^{\mathcal{B}}(\alpha) \equiv \mathrm{div}_P(\mathcal{B}^\sharp(\alpha)). \tag{6.18}$$

  On flat Euclidean space this recovers the diffusion-Stein operator constructed in [136], where the formal replacement "$\frac{1}{2}\nabla u \mapsto g$" therein is precisely the generalisation from exact 1-forms to arbitrary 1-forms $\mathrm{d}f \mapsto \alpha$. In other words, the $\mathcal{B}$-Stein operator on functions is precisely the restriction to exact forms of the bracket-Stein operator on differential forms. This shows that geometrically, the "diffusion"-Stein operator is simply

the restriction of the canonical Stein operator to the vector fields generated by $\mathcal{B}^\sharp$, for some bracket $\mathcal{B}$. Importantly, this does not rely on the generator/diffusion approach [24], and avoids the awkward formal replacement "$\frac{1}{2}\nabla u \mapsto g$". Similarly, the "Langevin Stein operator" [137] requires a Riemannian metric to construct the generator of the overdamped Langevin dynamics, $\mathcal{L}(f) = \frac{1}{2}\nabla \log p_\infty(f) + \frac{1}{2}\Delta f$ as well as the formal replacement $\frac{1}{2}\nabla f \mapsto X$, while its construction via the "product approach" (see (21) [154]) relies on the product structure of Euclidean space. Yet, as discussed above, none of these approaches are necessary: the smoothness and positivity of the target distribution canonically induce a Stein operator with the desired properties on general manifolds. We will construct diffusion Stein operators using the generator approach in (6.19).

As we noted, the canonical Stein operator naturally combines vector fields with the target measure to construct mean-zero functions, without relying on any extra structure on the sample space. One can of course also construct Stein operators on functions using the generator approach [24], which relies on the generator $\mathcal{L}_P$ of a P-preserving diffusion, $\mathcal{T}_{\mathcal{L}_P} \equiv \mathcal{L}_P$, since for f in the domain of $\mathcal{L}_P$

$$\int \mathcal{T}_P f P \int \mathcal{L}_P f P = \int f \mathcal{L}_P^* P = 0.$$

From our completeness result, we know that the measure-preserving generators are parametrised by the noise fields (Y_i), and antisymmetric brackets through their curl vector field ϕ_P:

$$\mathcal{T}_{\mathcal{L}_P} f = \phi_P(f) + \tfrac{1}{2}\mathrm{div}_P(Y_i)Y_i f + \tfrac{1}{2}Y_i Y_i f, \tag{6.19}$$

up to additional non-zero elements of the (first) homology group of P. Hence any Stein operator on a manifold constructed from the generator of a measure-preserving diffusion must have this form (at least locally), and we shall refer to $\mathcal{T}_{\mathcal{L}_P}$ as the **diffusion Stein operator**, which acts on functions in the domain of the generator.[3] The diffusion Stein operator arises from the $\mathcal{B}$-Stein operator (6.17) by choosing the symmetric part of $\mathcal{B}$ to be non-negative definite, which is equivalent to demanding that it can be expressed as the canonical symmetric bracket (recall §4.4.5) associated to a set of vector fields Y_i (see (6.2) [176]).

It is also possible to extend to manifolds the key result concerning the "density approach" on $\mathcal{X} = [a, b]$, that is theorem 2.1 [218], to show the canonical Stein operator characterises the target probability distribution (5.6). Indeed, as mentioned above, when $\mathcal{X}$ is connected and oriented, there is an isomorphism between the top compactly supported de Rham cohomology group and $\mathbb{R}$ given by integration, $\omega \mapsto \int_{\mathcal{X}} \omega$, so in particular any compactly supported smooth

[3] The diffusion Stein operator (6.19) may be sometimes extended to an operator on closed 1-forms by corollary 1 [207].

positive measure whose integral vanishes is exact. Thus, if $\mathcal{X}$ is compact and $C \subset \mathcal{X}$ is any closed (or open) set, fix $f_C \equiv \mathbb{I}_C - P(C)1$ where $\mathbb{I}_C$ is the indicator function at C. Then $\int f_C P = P(C) - P(C) = 0$, so $f_C P = \mathrm{d}\alpha_C$ for some α_C. Let $X_C \equiv P^\sharp(\alpha_C) \in \mathfrak{X}_P$ be the associated vector field, so $\mathrm{div}_P(X_C) = P^\sharp \mathrm{d}\alpha_C = f_C$. Thus $\int \mathrm{div}_P(X_C)Q = \int f_C Q = Q(C) - P(C)Q(\mathcal{X})$, and it follows by lemma 7.1.2 [50] that $\int \mathcal{T}_P(X)Q = 0$ for all $X \in \mathfrak{X}_P$ implies $Q = P$ (up to normalisation). Indeed in [218], $\mathcal{X} = [a, b]$, and the authors essentially simply use the fact that $\mathrm{d}\left(\int_a^z f_C(u)p(u)\mathrm{d}u\right) = f_C(z)p(z)\mathrm{d}z$ to explicitly construct $\alpha_C(z) \equiv \int_a^z f_C(u)p(u)\mathrm{d}u$ (and thus $X_C(z) = \frac{1}{p(z)} \int_a^z f_C P \partial_z$). From above we note that when $\mathcal{X}$ is a general manifold (perhaps with boundary), it suffices that for any open set C the top form $f_C P$ is exact (which holds for example on any non-compact, connected, orientable manifold), in which case $X_C \equiv P^\sharp(\alpha_C)$, where $f_C P = \mathrm{d}\alpha_C$, belongs to $\mathfrak{X}_P$ since $\int \mathrm{d}\alpha_C = \int f_C P = 0$, and the above derivation may be replicated.[4] Hence, again, we observe that we do not need to rely on generators to generalise the univariate case: once the appropriate geometric formulation of Stein operators has been introduced, many key statistical results may be obtained directly from very general results in differential geometry, which also allows us to better understand the influence of the topology of the sample space.

Finally, let us generalise two theorems in [224] (and provide much shorter proofs) which we will use to generalise KSD to manifolds in §A.7.1. The first is the key result of SVGD describing the rate of change of the KL statistical divergence between a 1-parameter family of measures Q_t satisfying the the continuity equation:

Theorem 49. *Let $Q_t = q_t \mu_\mathcal{X}$, $P = p_\infty \mu_\mathcal{X}$, where q_t satisfies $\frac{\partial q_t}{\partial t} + div(q_t Y) = 0$ for some complete vector field Y. Then*

$$-\frac{\mathrm{d}}{\mathrm{d}t} KL(Q_t \| P) = \int div_P(Y) Q_t.$$

Proof. Since Y is complete $\Phi_t^Y(\mathcal{X}) = \mathcal{X}$ for any t. Then

$$\begin{aligned}
-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(Q_t\|P) &= -\frac{\mathrm{d}}{\mathrm{d}t}\int_\mathcal{X} \log\left(\frac{q_t}{p_\infty}\right) Q_t = -\frac{\mathrm{d}}{\mathrm{d}t}\int_\mathcal{X} \log\left(\frac{q_t}{p_\infty}\right) q_t \mu_\mathcal{X} = -\frac{\mathrm{d}}{\mathrm{d}t}\int_\mathcal{X} \log\left(\frac{q_t}{p_\infty}\right) q_t \mu_\mathcal{X} \\
&= -\int_\mathcal{X} \left(\frac{\partial \log\left(\frac{q_t}{p_\infty}\right) q_t}{\partial t} + \mathrm{div}_{\mu_\mathcal{X}}\left(\log\left(\frac{q_t}{p_\infty}\right) q_t Y\right)\right) \mu_\mathcal{X} \\
&= -\int_\mathcal{X} \left(q_t \frac{\partial \log\left(\frac{q_t}{p_\infty}\right)}{\partial t} + q_t Y\left(\log\left(\frac{q_t}{p_\infty}\right)\right)\right) \mu_\mathcal{X}
\end{aligned}$$

[4] Note that the statement that a smooth measure Q is exact, is equivalent to the statement that its density f with respect to a smooth positive measure P, $Q = fP$, can be expressed as the P-divergence of a vector field.

$$
\begin{aligned}
&= -\int_{\mathcal{X}} \left(\frac{\partial \log q_t}{\partial t} + Y\left(\log\left(\frac{q_t}{p_\infty}\right)\right)\right) Q_t \\
&= -\int_{\mathcal{X}} \left(-\mathrm{div}_{\mu_{\mathcal{X}}}(q_t Y) + Y\left(\log q_t - \log p_\infty\right)\right) Q_t \\
&= -\int_{\mathcal{X}} \left(-\mathrm{div}_{\mu_{\mathcal{X}}}(Y) - Y\left(\log p_\infty\right)\right) Q_t = \int_{\mathcal{X}} \mathrm{div}_P\left(Y\right) Q_t.
\end{aligned}
$$

■

The second generalisation is the following very useful proposition which shows that if a vector space $\mathcal{H}$ intersects the Casimirs of a bracket trivially, then $\mathcal{H} \cong \mathfrak{X}_{\mathcal{B}}|_{\mathcal{H}} \equiv \{i_{\mathrm{d}h}\mathcal{B} : h \in \mathcal{H}\}$ in the category of vector spaces (we recover [224] when $\mathcal{B} = G^{-1}$ and $\mathcal{H}$ is a RKHS containing only 0 as constant function).

Proposition 50. *Let $\mathcal{B}$ be a bracket and $\mathcal{H}$ a vector space of functions on $\mathcal{X}$ s.t., Casimir$(\mathcal{B}) \bigcap \mathcal{H} = \{0\}$. Then the map $f \mapsto X_f$ is a vector space isomorphism which we can use to transfer structures on $\mathcal{H}$ to $\mathfrak{X}_{\mathcal{B}}|_{\mathcal{H}}$. For example if $\mathcal{H}$ is an inner product space then so is $\mathfrak{X}_{\mathcal{B}}|_{\mathcal{H}}$ with $\langle X_f, X_h\rangle \equiv \langle f, h\rangle$. In particular, if $\mathcal{B}$ is non-degenerate this holds iff $\mathcal{H}$ is a space of functions for which 0 is the only locally constant function.*

Proof. Since Casimir$(\mathcal{B}) \bigcap \mathcal{H} = \{0\}$ the map $h \mapsto X_h$ is injective, and thus a bijection onto its image. Since the map is $\mathbb{R}$-linear this is a vector space isomorphism. It follows that $\langle X_f, X_h\rangle \equiv \langle f, h\rangle$ is a well-defined inner product. In particular if $\mathcal{B}$ is non-degenerate the only Casimirs are the locally constant functions. ■

We mention that if P is not positive globally but vanishes outside a submanifold $\mathcal{S}_P$, then setting $P \equiv f\mu_{\mathcal{X}}$ for a smooth positive reference measure $\mu_{\mathcal{X}}$, and a non-negative function f with $f > 0$ on $\mathcal{S}_P$, we have

$$
\begin{aligned}
\mathcal{L}_X P &= X(f)\mu_{\mathcal{X}} + f\mathrm{div}_{\mu_{\mathcal{X}}}(X)\mu_{\mathcal{X}} = \left(X(f)\mu_{\mathcal{X}} + f\mathrm{div}_{\mu_{\mathcal{X}}}(X)\mu_{\mathcal{X}}\right)\mathbb{I}_{\mathcal{S}_P} \\
&= \left(X(f) + f\left(\mathrm{div}_P(X) - X(\log f)\right)\right)\mu_{\mathcal{X}}\mathbb{I}_{\mathcal{S}_P} = \mathrm{div}_P(X)P\mathbb{I}_{\mathcal{S}_P},
\end{aligned}
$$

Hence the Stein operator of P may be defined as $\mathrm{div}_P(X)$ on $\mathcal{S}_P$, and 0 otherwise.

Next, we will use $\mathcal{B}$-Stein operators to generalise diffusion score-matching to manifolds.

6.10 Bracket-Induced Stein Discrepancies

The **$\mathcal{W}$-Stein discrepancy** of the canonical Stein operator associated to a family of vector fields $\mathcal{W} \subset \mathfrak{X}_P$ is defined as

$$\mathrm{SD}_{\mathcal{W}}(P\|\mathbb{P}_\theta) \equiv \sup_{f\in\mathcal{T}_{\mathbb{P}_\theta}(\mathcal{W})} \left|\int fP - \int f\mathbb{P}_\theta\right| = \sup_{X\in\mathcal{W}} \left|\int \mathcal{T}_{\mathbb{P}_\theta}(X)P\right|$$

where here $\mathbb{P}_\theta$ belongs to a smooth statistical model $\mathcal{P}_\Theta$ as above. As in §6.9 we can use a bracket and let $\mathcal{W} = \{i_\alpha\mathcal{B} : \alpha \in \mathcal{H}\}$ for some set of 1-forms $\mathcal{H} \subset \Omega^1_P$. The **$\mathcal{B}$-Stein discrepancy over $\mathcal{H}$** is then defined by

$$\mathrm{SD}^{\mathcal{B}}_{\mathcal{H}}(P\|\mathbb{P}_\theta) \equiv \sup_{\alpha\in\mathcal{H}} \left|\int \mathcal{T}^{\mathcal{B}}_{\mathbb{P}_\theta}(\alpha)P\right|.$$

In general we have the following useful representation of the $\mathcal{B}$-SD in terms of the pairing (4.8) generated by the measure:

Theorem 51. *Suppose $\mathbb{P}_\theta = e^{-H_\theta}P_{\mathcal{X}}$ and $P = e^{-H}P_{\mathcal{X}}$ for some smooth reference measure $P_{\mathcal{X}}$. If $\mathcal{H} \subset \Omega^1_P$, then*

$$\mathrm{SD}^{\mathcal{B}}_{\mathcal{H}}(P\|\mathbb{P}_\theta) = \sup_{\alpha\in\mathcal{H}} \left|\left\langle \alpha, X^{\mathcal{B}^*}_{H-H_\theta}\right\rangle_P\right|, \tag{6.20}$$

where $\mathcal{B}^*$ is the dual bracket (see 4.3.1.1). In particular if $\mathcal{H}$ is a dense subspace of the unit ball of some normed vector space $(V, \|\cdot\|_V)$, with dual norm denoted $\|\cdot\|^*_V$, and $\left\langle \cdot, X^{\mathcal{B}^*}_{H-H_\theta}\right\rangle_P \in V^*$ for any $\theta \in \Theta$:

$$\mathrm{SD}^{\mathcal{B}}_{\mathcal{H}}(P\|\mathbb{P}_\theta) = \sup_{\alpha\in\mathcal{H}} \left|\left\langle \alpha, X^{\mathcal{B}^*}_{H-H_\theta}\right\rangle_P\right| = \left\|\left\langle \cdot, X^{\mathcal{B}^*}_{H-H_\theta}\right\rangle_P\right\|^*_V$$

Proof. Since $\alpha \in \Omega^1_P$

$$\begin{aligned}
\int \mathcal{T}^{\mathcal{B}}_{\mathbb{P}_\theta}(\alpha)P &= \int \left(\mathcal{T}^{\mathcal{B}}_{\mathbb{P}_\theta}(\alpha) - \mathcal{T}^{\mathcal{B}}_{P}(\alpha)\right)P = \int \left(\mathrm{div}_{\mathbb{P}_\theta}(i_\alpha\mathcal{B}) - \mathrm{div}_P(i_\alpha\mathcal{B})\right)P \\
&= \int \left(\mathcal{B}(\alpha, \mathrm{d}H) - \mathcal{B}(\alpha, \mathrm{d}H_\theta)\right) = \int \left\langle \mathcal{B}^\sharp(\alpha), \mathrm{d}H - \mathrm{d}H_\theta\right\rangle P \\
&= \int \left\langle \alpha, \mathcal{B}^{*\sharp}(\mathrm{d}H - \mathrm{d}H_\theta)\right\rangle P,
\end{aligned}$$

from which the result follows. ∎

In particular if $\mathcal{X}$ is Riemannian and $P = e^{-H}\mathrm{vol}$, choosing $V \equiv L^2(T^*\mathcal{X}, P)$, the Hilbert space of square P-integrable 1-forms, $(\alpha,\alpha)_P < \infty$, where the inner product is defined by (see (4.9))

$$(\alpha,\gamma)_P \equiv (e^{-H}\alpha, \gamma)_{\mathrm{vol}} = \int e^{-H}\alpha \wedge \star\gamma = \int \langle \alpha, \gamma\rangle^{-1} P,$$

we obtain the generalisation of the score matching (see §A.7.1 for the generalisation of KSD):

Corollary 52. *We define* $\mathcal{B}$***-score-matching*** $\mathrm{SM}_{\mathcal{B}}$*, between probability measures that are absolutely continuous with respect to the Riemannian measure vol, as the* $\mathcal{W}$*-canonical* $\mathcal{B}$*-Stein discrepancy with* $\mathcal{W} \equiv \{i_\alpha \mathcal{B} : (\alpha, \alpha)_P \leq 1 \text{ and } \alpha \in \Omega^1_P\}$*. Whenever* $\mathcal{W}$ *is a dense subset of the unit ball in* $L^2(T^*\mathcal{X}, P)$*, the score matching discrepancy from* $P \propto e^{-H}\mathrm{vol}$ *to distributions in a statistical model* $\{\mathbb{P}_\theta = e^{-H_\theta}\mathrm{vol} : \theta \in \Theta\}$

$$\mathrm{SM}_{\mathcal{B}}(P\|\mathbb{P}_\theta) = \int \|X^{\mathcal{B}^*}_{H-H_\theta}\|^2 P = \mathbb{E}_P\left[\|X^{\mathcal{B}^*}_{H-H_\theta}\|^2\right]. \tag{6.21}$$

$\mathcal{B}$*-score matching is a statistical divergence, that is* $\mathrm{SM}_{\mathcal{B}}(P\|\mathbb{P}_\theta) = 0$ *iff* $H_\theta = H$ *vol-a.e., whenever* $\mathrm{Casimirs}(\mathcal{B}^*) \bigcap \{H_\theta - H : \theta \in \Theta\} = \{\text{constants}\}$ *vol-a.e.. The latter always holds when* $\mathcal{B}^*$ *is non-degenerate. Moreover, using the* $\mathcal{B}$*-integration by parts, we can generalise Hyvarinen's trick and obtain*

$$\mathrm{SM}_{\mathcal{B}}(P\|\mathbb{P}_\theta) = \underbrace{\int \left(\|X^{\mathcal{B}^*}_{H_\theta}\|^2 + 2\nabla \cdot \left(\sharp\mathcal{B}^{*\sharp^\top} X^{\mathcal{B}^*}_{H_\theta}\right)\right) P}_{\text{depends on } P \text{ only through integration}} + \underbrace{\int \|X^{\mathcal{B}^*}_{H}\|^2 P}_{\theta\text{-independent}}, \tag{6.22}$$

where $\mathcal{B}^{*\sharp^\top} : \mathfrak{X}(\mathcal{X}) \to \Omega^1(\mathcal{X})$ *is the transpose of* $\mathcal{B}^{*\sharp}$ *with respect to the Riemannian metrics.*

Proof. Starting from (6.20)

$$\begin{aligned}
&\sup_{(\alpha,\alpha)_P \leq 1} \left|\left\langle \alpha, X^{\mathcal{B}^*}_{H-H_\theta}\right\rangle_P\right| = \sup_{(\alpha,\alpha)_P \leq 1} \left|\left\langle p_\infty \alpha, X^{\mathcal{B}^*}_{H-H_\theta}\right\rangle_{\mathrm{vol}}\right| \\
&= \sup_{(\alpha,\alpha)_P \leq 1} \left|\left(\alpha, \flat X^{\mathcal{B}^*}_{H-H_\theta}\right)_P\right| = \|\flat X^{\mathcal{B}^*}_{H-H_\theta}\|_P.
\end{aligned}$$

Moreover $\mathrm{SM}_{\mathcal{B}}(P\|\mathbb{P}_\theta) = 0$ iff $\|X^{\mathcal{B}^*}_{H-H_\theta}\| = 0$ vol-a.e. iff $X^{\mathcal{B}^*}_{H-H_\theta} = 0$ vol-a.e. iff $H - H_\theta \in \mathrm{Casimirs}(\mathcal{B}^*)$ vol-a.e.. Hence $H - H_\theta$ is a constant vol-a.e., so $H = H_\theta$ vol-a.e. since the probability measures $P, \mathbb{P}_\theta$ must integrate to 1. The last equality follows from

$$\begin{aligned}
\int \left\langle X^{\mathcal{B}^*}_H, X^{\mathcal{B}^*}_{H_\theta}\right\rangle P &= \int \left\langle \mathcal{B}^{*\sharp}(\mathrm{d}H), X^{\mathcal{B}^*}_{H_\theta}\right\rangle P = \int \left\langle \mathrm{d}H, \mathcal{B}^{*\sharp^\top} X^{\mathcal{B}^*}_{H_\theta}\right\rangle^{-1} P \\
&= -\int \left\langle \mathrm{d}e^{-H}, \mathcal{B}^{*\sharp^\top} X^{\mathcal{B}^*}_{H_\theta}\right\rangle^{-1} \mathrm{vol} \\
&= \int \nabla \cdot \left(\sharp\mathcal{B}^{*\sharp^\top} X^{\mathcal{B}^*}_{H_\theta}\right) P,
\end{aligned}$$

where the last line follows from $\mathcal{B}$-integration by parts 7. ∎

When $\mathcal{B}$ is Riemannian we recover the Riemannian score matching

$$\mathrm{SM}_{G^{-1}}(P\|\mathbb{P}_\theta) = \int \|\nabla \log p_\theta - \nabla \log p_\infty\|^2 P,$$

while in Euclidean space (6.21) yields the diffusion score matching. In (6.21) we also observe that, in the spirit of [240], the $\mathcal{B}^*$-Hamiltonian vector field $X_H^{\mathcal{B}^*}$ generalises the score $\nabla \log p_\infty$ of the target density: $\mathcal{B}$-score matching matches the $\mathcal{B}$-Hamiltonian fields of the input measures. The representation (6.22) has the important property that its first term does not depend on θ, and may be ignored during the process of optimising $\theta \mapsto \mathrm{SM}_{\mathcal{B}}(P\|\mathbb{P}_\theta)$, while the other terms only depend on P through integration and may thus be estimated using a U-statistics.

The reminder of this chapter contains the proofs of the results stated beforehand, as well as additional details on the Stein discrepancy minimum estimators.

6.11 Derivations on Diffusion Stein Discrepancies

By definition for $f \in \Gamma(\mathcal{X}, \mathbb{R}^d)$ and $A \in \Gamma(\mathcal{X}, \mathbb{R}^{d\times d})$

$$\begin{aligned}\mathcal{S}_p[f] &= \frac{1}{p}\nabla \cdot (pmf) = m^\top \nabla \log p \cdot f + \nabla \cdot (mf),\\ \mathcal{S}_p[A] &= \frac{1}{p}\nabla \cdot (pmA) = m^\top \nabla \log p \cdot A + \nabla \cdot (mA)\end{aligned}$$

which are operators $\Gamma(\mathcal{X}, \mathbb{R}^d) \to \Gamma(\mathcal{X}, \mathbb{R})$ and $\Gamma(\mathcal{X}, \mathbb{R}^{d\times d}) \to \Gamma(\mathcal{X}, \mathbb{R}^d)$ respectively.

Proposition 53. *Let $\mathcal{X}$ be an open (connected) subset of $\mathbb{R}^d$, m is continuously differentiable, and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$ is the matrix kernel of $\mathcal{H}^d$. Suppose for any $j \in [1, d]$, $K, \partial_{1j}\partial_{2j}K$ are separately continuous and locally bounded. Then for any $f \in \mathcal{H}^d$*

$$\mathcal{S}_p[f](x) = \langle \mathcal{S}_p^1[K]|_x, f\rangle_{\mathcal{H}^d}$$

Proof. Note that technically the kernel K of $\mathcal{H}^d$ takes value in the set of (bounded) linear operators on $\mathbb{R}^d$, and we view these linear operators as matrices by defining the components $(K(x,y))_{ji} \equiv e_j \cdot K(x,y)e_i$, where (e_l) is the canonical basis of $\mathbb{R}^d$. For any $f \in \mathcal{H}^d$

$$\begin{aligned}\langle f(x), m(x)^\top \nabla \log p(x)\rangle_{\mathbb{R}^d} &= \langle f, K(\cdot, x)m(x)^\top \nabla \log p(x)\rangle_{\mathcal{H}^d}\\ &= \langle f, K_x^\top m(x)^\top \nabla \log p(x)\rangle_{\mathcal{H}^d}\\ &= \langle f, m(x)^\top \nabla \log p(x) \cdot K_x\rangle_{\mathcal{H}^d}.\end{aligned}$$

Moreover, under these assumptions the RKHS $\mathcal{H}^d$ is continuously embedded in the topological space $C^1(\mathcal{X}, \mathbb{R}^d)$, so its elements are continuously differentiable. Then for any $f \in \mathcal{H}^d$, by theorem 2.11 [263]

$$\langle f, \partial_{2^j} K(\cdot, x) e_r \rangle_{\mathcal{H}^d} = \langle e_r, \partial_j f|_x \rangle_{\mathbb{R}^d} = \partial_j f_r|_x.$$

Hence

$$\begin{aligned}
\langle f, \nabla \cdot (mK)|_x \rangle_{\mathcal{H}^d} &= \langle f, \partial_{1^j}(m_{jr} K_{ri})|_x e_i \rangle_{\mathcal{H}^d} = \langle f, \partial_j m_{jr}|_x K_{ri}(x, \cdot) e_i + m_{jr}(x) \partial_{1^j} K_{ri}|_x e_i \rangle_{\mathcal{H}^d} \\
&= \partial_j m_{jr}|_x \langle f, K_{ir}(\cdot, x) e_i \rangle_{\mathcal{H}^d} + m_{jr}(x) \langle f, \partial_{1^j} K_{ri}(x, \cdot) e_i \rangle_{\mathcal{H}^d} \\
&= \partial_j m_{jr}|_x \langle f, K(\cdot, x) e_r \rangle_{\mathcal{H}^d} + m_{jr}(x) \langle f, \partial_{2^j} K_{ir}(\cdot, x) e_i \rangle_{\mathcal{H}^d} \\
&= \partial_j m_{jr}|_x \langle f, K(\cdot, x) e_r \rangle_{\mathcal{H}^d} + m_{jr}(x) \langle f, \partial_{2^j} K(\cdot, x) e_r \rangle_{\mathcal{H}^d} \\
&= \partial_j m_{jr}|_x f_r(x) + m_{jr}(x) \partial_j f_r|_x \\
&= \langle \nabla \cdot m, f(x) \rangle_{\mathbb{R}^d} + \operatorname{tr}[m(x) \nabla_x f] = \nabla_x \cdot (mf).
\end{aligned}$$

Therefore, we conclude that $\mathcal{S}_p[f](x) = \langle \mathcal{S}_p^1 K_x, f \rangle_{\mathcal{H}^d}$ where $\mathcal{S}_p^1 K_x \equiv \mathcal{S}_p^1[K]|_x$ means applying $\mathcal{S}_p$ to the first entry of K and evaluate it x, so informally $\mathcal{S}_p^1[K]|_x : y \mapsto \frac{1}{p} \nabla_x \cdot (p(x) m(x) K(x, y))$. ■

6.11.1 Diffusion Kernel Stein Discrepancies

Proposition 54. *Suppose $\mathcal{S}_p[f](x) = \langle \mathcal{S}_p^1[K]|_x, f \rangle_{\mathcal{H}^d}$ for any $f \in \mathcal{H}^d$. Let m and K be C^2, and $x \mapsto \mathcal{S}_p K_x$ be $\mathbb{Q}$-Bochner integrable. Then*

$$\mathrm{DKSD}_{K,m}(\mathbb{Q}, \mathbb{P})^2 = \int_{\mathcal{X}} \int_{\mathcal{X}} \mathcal{S}_p^2 \mathcal{S}_p^1 K(x, y) \mathrm{d}\mathbb{Q}(x) \mathrm{d}\mathbb{Q}(y).$$

Proof. Let us identify $\mathcal{H}_1 \otimes \mathcal{H}_2 \cong L(\mathcal{H}_1 \times \mathcal{H}_2, \mathbb{R}) \cong L(\mathcal{H}_2, \mathcal{H}_1)$ with $(v_1 \otimes v_2) \sim v_1 \langle v_2, \cdot \rangle_{\mathcal{H}_2}$ (since $\mathcal{H}_2 \cong \mathcal{H}_2^*$), so that $(v_1 \otimes v_2) u_2 \equiv v_1 \langle v_2, u_2 \rangle_{\mathcal{H}_2}$ (here $L(V, W)$ is the space of linear maps from V to W). Then

$$\langle u_1 \otimes u_2, v_1 \otimes v_2 \rangle_{HS} \equiv \langle u_1, v_1 \rangle_{\mathcal{H}_1} \langle u_2, v_2 \rangle_{\mathcal{H}_2} = \langle u_1, (v_1 \otimes u_2) v_2 \rangle_{\mathcal{H}_1}.$$

For simplicity we will write $\mathcal{S}_p K_x \equiv \mathcal{S}_p^1[K]|_x$. Using the fact $x \mapsto \mathcal{S}_p K_x$ is $\mathbb{Q}$-Bochner integrable, then by Cauchy-Schwartz $x \mapsto \langle h, \mathcal{S}_p K_x \rangle_{\mathcal{H}^d}$ is $\mathbb{Q}$-integrable. Then

$$\mathrm{DKSD}_{K,m}(\mathbb{Q}, \mathbb{P})^2 = \sup_{\substack{h \in \mathcal{H}^d \\ \|h\| \leq 1}} \left\langle \int_{\mathcal{X}} \mathcal{S}_p[h](x) \mathrm{d}\mathbb{Q}(x), \int_{\mathcal{X}} \mathcal{S}_p[h](y) \mathrm{d}\mathbb{Q}(y) \right\rangle_{\mathbb{R}}$$

$$
\begin{aligned}
&= \sup_{\substack{h\in\mathcal{H}^d\\ \|h\|\leq 1}} \int_{\mathcal{X}} \langle h, \mathcal{S}_p K_x\rangle_{\mathcal{H}^d} \mathrm{d}\mathbb{Q}(x) \int_{\mathcal{X}} \langle h, \mathcal{S}_p K_y\rangle_{\mathcal{H}^d} \mathrm{d}\mathbb{Q}(y) \\
&= \sup_{\substack{h\in\mathcal{H}^d\\ \|h\|\leq 1}} \int_{\mathcal{X}}\int_{\mathcal{X}} \langle h, \mathcal{S}_p K_x\rangle_{\mathcal{H}^d} \langle h, \mathcal{S}_p K_y\rangle_{\mathcal{H}^d} \mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y) \\
&= \sup_{\substack{h\in\mathcal{H}^d\\ \|h\|\leq 1}} \int_{\mathcal{X}}\int_{\mathcal{X}} \langle h, \mathcal{S}_p K_x \otimes \mathcal{S}_p K_y h\rangle_{\mathcal{H}^d} \mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y) \\
&= \sup_{\substack{h\in\mathcal{H}^d\\ \|h\|\leq 1}} \int_{\mathcal{X}}\int_{\mathcal{X}} \langle h\otimes h, \mathcal{S}_p K_x \otimes \mathcal{S}_p K_y\rangle_{HS} \mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y).
\end{aligned}
$$

Moreover $\int_{\mathcal{X}} \|\mathcal{S}_p K_x \otimes \mathcal{S}_p K_y\|_{HS} \mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y) < \infty$, since

$$
\begin{aligned}
&\int_{\mathcal{X}} \|\mathcal{S}_p K_x \otimes \mathcal{S}_p K_y\|_{HS} \mathrm{d}\mathbb{Q}(x) \otimes \mathrm{d}\mathbb{Q}(y) \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} \sqrt{\langle \mathcal{S}_p K_x, \mathcal{S}_p K_x\rangle_{\mathcal{H}^d} \langle \mathcal{S}_p K_y, \mathcal{S}_p K_y\rangle_{\mathcal{H}^d}} \mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y) \\
&= \left(\int_{\mathcal{X}} \sqrt{\langle \mathcal{S}_p K_x, \mathcal{S}_p K_x\rangle_{\mathcal{H}^d}} \mathrm{d}\mathbb{Q}(x)\right)^2 \\
&= \left(\int_{\mathcal{X}} \|\mathcal{S}_p K_x\|_{\mathcal{H}^d} \mathrm{d}\mathbb{Q}(x)\right)^2 < \infty
\end{aligned}
$$

since by assumption $x \mapsto \mathcal{S}_p K_x$ is $\mathbb{Q}$-Bochner integrable. Thus

$$
\begin{aligned}
\mathrm{DKSD}_{K,m}(\mathbb{Q},\mathbb{P})^2 &= \sup_{\substack{h\in\mathcal{H}^d\\ \|h\|\leq 1}} \left\langle h\otimes h, \int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{S}_p K_x \otimes \mathcal{S}_p K_y \mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y)\right\rangle_{HS} \\
&= \left\| \int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{S}_p K_x \otimes \mathcal{S}_p K_y \mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y)\right\|_{HS} \\
&= \left\| \int_{\mathcal{X}} \mathcal{S}_p K_x \mathrm{d}\mathbb{Q}(x) \otimes \int_{\mathcal{X}} \mathcal{S}_p K_y \mathrm{d}\mathbb{Q}(y)\right\|_{HS} \\
&= \left\| \int_{\mathcal{X}} \mathcal{S}_p K_x \mathrm{d}\mathbb{Q}(x)\right\|^2_{\mathcal{H}^d} \\
&= \left\langle \int_{\mathcal{X}} \mathcal{S}_p K_x \mathrm{d}\mathbb{Q}(x), \int_{\mathcal{X}} \mathcal{S}_p K_y \mathbb{Q}(\mathrm{d}y)\right\rangle_{\mathcal{H}^d} \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} \langle \mathcal{S}_p K_x, \mathcal{S}_p K_y\rangle_{\mathcal{H}^d} \mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y) \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{S}_p^2 \mathcal{S}_p^1 K(x,y) \mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y).
\end{aligned}
$$

To show the penultimate equality (exchange integral and inner product), we use the fact $\mathcal{S}_p K_x$ is $\mathbb{Q}$-Bochner integrable, and that the operator $W : f \mapsto \langle f, \int_{\mathcal{X}} \mathcal{S}_p K_y \mathbb{Q}(\mathrm{d}y)\rangle_{\mathcal{H}^d}$ is bounded, from

which it follows that

$$\begin{aligned}\left\langle \int_{\mathcal{X}} \mathcal{S}_p K_x \mathrm{d}\mathbb{Q}(x), \int_{\mathcal{X}} \mathcal{S}_p K_y \mathbb{Q}(\mathrm{d}y) \right\rangle_{\mathcal{H}^d} &= W\left[\int_{\mathcal{X}} \mathcal{S}_p K_x \mathrm{d}\mathbb{Q}(x)\right] = \int_{\mathcal{X}} W\left[\mathcal{S}_p K_x \mathrm{d}\mathbb{Q}(x)\right] \\ &= \int_{\mathcal{X}} \left\langle \mathcal{S}_p K_x, \int_{\mathcal{X}} \mathcal{S}_p K_y \mathrm{d}\mathbb{Q}(y) \right\rangle_{\mathcal{H}^d} \mathrm{d}\mathbb{Q}(x) \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} \langle \mathcal{S}_p K_x, \mathcal{S}_p K_y \rangle_{\mathcal{H}^d} \mathrm{d}\mathbb{Q}(x) \mathrm{d}\mathbb{Q}(y)\end{aligned}$$

Hence $\mathrm{DKSD}_{K,m}(\mathbb{Q},\mathbb{P})^2 = \int_{\mathcal{X}} \int_{\mathcal{X}} \mathcal{S}_p^2 \mathcal{S}_p^1 K(x,y) \mathrm{d}\mathbb{Q}(x) \mathrm{d}\mathbb{Q}(y)$.

Note that from this proof we have

$$k^0(x,y) \equiv \mathcal{S}_p^2 \mathcal{S}_p^1 K(x,y) = \langle \mathcal{S}_p K_x, \mathcal{S}_p K_y \rangle_{\mathcal{H}^d},$$

which shows the map $\phi : \mathcal{X} \to \mathcal{H}^d$, $\phi(x) \equiv \mathcal{S}_p^1[K]|_x$ is a feature map (more precisely it is dual to the feature map) for the scalar reproducing kernel k^0, and its RKHS consists of functions $g(\cdot) = \langle \phi(\cdot), f \rangle_{\mathcal{H}^d}$ for $f \in \mathcal{H}^d$ [262]. ■

6.11.2 Diffusion Kernel Stein Discrepancies as Statistical Divergences

In this section, we prove that DKSD is a statistical divergence and provide sufficent conditions on the matrix-valued kernel.

6.11.2.1 Proof of 36: DKSD as statistical divergence

By Stoke's theorem $\int_{\mathcal{X}} \mathcal{S}_q[v] \mathrm{d}\mathbb{Q} = \int_{\mathcal{X}} \nabla \cdot (qmv) \mathrm{d}x = 0$, thus $\int_{\mathcal{X}} \mathcal{S}_p[v] \mathrm{d}\mathbb{Q} = \int_{\mathcal{X}} (\mathcal{S}_p[v] - \mathcal{S}_q[v]) \mathrm{d}\mathbb{Q} = \int_{\mathcal{X}} (s_p - s_q) \cdot v \mathrm{d}\mathbb{Q}$, and by assumption $\int_{\mathcal{X}} \mathcal{S}_q[K] \mathrm{d}\mathbb{Q} = \int_{\mathcal{X}} \nabla \cdot (qmK) \mathrm{d}x = 0$. Moreover, with $s_p = m^\top \nabla \log p$, and $\delta_{p,q} \equiv s_p - s_q$. Hence

$$\begin{aligned}\mathrm{DKSD}_{K,m}(\mathbb{Q},\mathbb{P})^2 &= \int_{\mathcal{X}} \int_{\mathcal{X}} \mathcal{S}_p^2 \left[\mathcal{S}_p^1 K(x,y)\right] \mathrm{d}\mathbb{Q}(y) \mathrm{d}\mathbb{Q}(x) \\ &= \int_{\mathcal{X}} \int_{\mathcal{X}} (s_p(y) - s_p(y)) \cdot \left[\mathcal{S}_p^1 K(x,y)\right] \mathrm{d}\mathbb{Q}(y) \mathrm{d}\mathbb{Q}(x) \\ &= \int_{\mathcal{X}} (s_p(y) - s_p(y)) \mathrm{d}\mathbb{Q}(y) \cdot \int_{\mathcal{X}} \left[\mathcal{S}_p^1 K(x,y)\right] \mathrm{d}\mathbb{Q}(x) \\ &= \int_{\mathcal{X}} (s_p(y) - s_p(y)) \mathrm{d}\mathbb{Q}(y) \cdot \int_{\mathcal{X}} \left[\mathcal{S}_p^1 K(x,y) - \mathcal{S}_q^1 K(x,y)\right] \mathrm{d}\mathbb{Q}(x) \\ &= \int_{\mathcal{X}} (s_p(y) - s_p(y)) \mathrm{d}\mathbb{Q}(y) \cdot \int_{\mathcal{X}} \left[(s_p(x) - s_p(x)) \cdot K(x,y)\right] \mathrm{d}\mathbb{Q}(x)\end{aligned}$$

$$= \int_{\mathcal{X}} \int_{\mathcal{X}} q(x)\delta_{p,q}(x)^\top K(x,y)\delta_{p,q}(y)q(y)\mathrm{d}x\mathrm{d}y$$
$$= \int_{\mathcal{X}} \int_{\mathcal{X}} \mathrm{d}\mu^\top(x)K(x,y)\mathrm{d}\mu(y).$$

where $\mu(\mathrm{d}x) \equiv q(x)\delta_{p,q}(x)\mathrm{d}x$, which is a finite measure by assumption. If $\mathcal{S}(q,p) = 0$, then since K is IPD we have $q\delta_{p,q} \equiv 0$, and since $q > 0$ and m is invertible we must have $\nabla \log p = \nabla \log q$ and thus $q = p$.

6.11.2.2 *Proof of 37: IPD matrix kernels*

Let μ be a finite signed vector measure. (i) If each k^i is IPD, then $\int \mathrm{d}\mu^\top K \mathrm{d}\mu = \int k^i(x,y)\mathrm{d}\mu_i(x)\mathrm{d}\mu_i(y) \geq 0$ with equality iff $\mu_i \equiv 0$ for all i. Conversely suppose $\int k^i(x,y)\mathrm{d}\mu_i(x)\mathrm{d}\mu_i(y) \geq 0$ with equality iff $\mu_i \equiv 0$ for all i . Suppose k^j is not IPD for some j, then there exists a finite non-zero signed measure ν s.t., $\int k^j \mathrm{d}\nu \otimes \mathrm{d}\nu \leq 0$, so if we define the vector measure $\mu_i \equiv \delta_{ij}\nu$, which is non-zero and finite, then $\int k^i(x,y)\mathrm{d}\mu_i(x)d\mu_i(y) \leq 0$ which contradicts the assumption. For (ii), we first diagonalise $B = R^\top DR$ where R is orthogonal and D diagonal with positive entries $\lambda_i > 0$. Then

$$\int \mathrm{d}\mu^\top K \mathrm{d}\mu = \int k\mathrm{d}\mu^\top R^\top DR\mathrm{d}\mu = \int k\left(R\mathrm{d}\mu\right)^\top D\left(R\mathrm{d}\mu\right) = \int k(x,y)\lambda_i \mathrm{d}\nu_i(x)\mathrm{d}\nu_i(y),$$

where $\nu \equiv R\mu$ is finite and non-zero, since μ is non-zero and R is invertible, thus maps non-zero vectors to non-zero vectors. Clearly if k is IPD then $\int \mathrm{d}\mu^\top K\mathrm{d}\mu \geq 0$ with equality iff $\nu_i \equiv 0$ for all i. Suppose K is IPD but k is not, then there exists finite non-zero signed measure ν for which $\int k\mathrm{d}\nu \otimes \mathrm{d}\nu \leq 0$, but then setting $\mu \equiv R^\top \xi$, with $\xi_i \equiv \delta_{ij}\nu$ which is finite and non-zero, implies $\int \mathrm{d}\mu^\top K\mathrm{d}\mu = \int k\mathrm{d}\xi^\top D\mathrm{d}\xi = \lambda_j \int k\mathrm{d}\nu \otimes \mathrm{d}\nu \leq 0$.

6.11.3 Diffusion Score Matching Derivations

6.11.3.1 *Proof of 38: Diffusion Score Matching*

Note that the Stein operator satisfies

$$\mathcal{S}_p[g] = \frac{\nabla \cdot (pmg)}{p} = \frac{\langle \nabla p, mg\rangle + p\nabla\cdot(mg)}{p} = \langle \nabla \log p, mg\rangle + \nabla\cdot(mg)$$
$$= \left\langle m^\top \nabla \log p, g\right\rangle + \nabla\cdot(mg).$$

Since $\int_{\mathcal{X}} \mathcal{S}_q[g]\mathrm{d}\mathbb{Q} = 0$, we have

$$\begin{aligned}
D(\mathbb{Q}\|\mathbb{P}) &= \sup_{g\in\mathcal{G}} \left| \int_{\mathcal{X}} \mathcal{S}_p[g](x)\mathbb{Q}(\mathrm{d}x) \right|^2 = \sup_{g\in\mathcal{G}} \left| \int_{\mathcal{X}} (\mathcal{S}_p[g](x) - \mathcal{S}_q[g](x))\mathbb{Q}(\mathrm{d}x) \right|^2 \\
&= \sup_{g\in\mathcal{G}} \left| \int_{\mathcal{X}} ((\nabla \log p - \nabla \log q)\cdot(mg))\,\mathrm{d}\mathbb{Q} \right|^2, \\
&= \sup_{g\in\mathcal{G}} \left| \left\langle m^\top (\nabla \log p - \nabla \log q), g \right\rangle_{L^2(\mathbb{Q})} \right|^2 \\
&= \left\| m^\top (\nabla \log p - \nabla \log q) \right\|^2_{L^2(\mathbb{Q})} = \int_{\mathcal{X}} \left\| m^\top (\nabla \log p - \nabla \log q) \right\|_2^2 \mathrm{d}\mathbb{Q},
\end{aligned}$$

where we have used the fact that $\mathcal{G}$ is dense in the unit ball of $L^2(\mathbb{Q})$ (since smooth functions with compact support are dense in $L^2(\mathbb{Q})$), and that the supremum over a dense subset of the continuous functional $F(\cdot) \equiv \left\langle m^\top (\nabla \log p - \nabla \log q), \cdot \right\rangle_{L^2(\mathbb{Q})}$ is equal to the supremum over the closure, $\sup_{\mathcal{G}} F = \sup_{\overline{\mathcal{G}}} F$. Suppose $D(\mathbb{Q}\|\mathbb{P}) = 0$. Then since $q > 0$ we must have $\left\| m^\top (\nabla \log p - \nabla \log q) \right\|_2^2 = 0$, i.e., $m^\top (\nabla \log p - \nabla \log q) = 0$, i.e., $\nabla(\log p - \log q) = 0$. Thus $\log(p/q) = c$, so $p = qe^c$ and integrating implies $c = 0$, so $D(\mathbb{Q}\|\mathbb{P}) = 0$ iff $\mathbb{Q} = \mathbb{P}$ a.e.. To obtain the estimator we will use the divergence theorem, which holds for example if $X, \nabla \cdot X \in L^1(\mathbb{R}^d)$ for $X = qmm^\top \nabla \log p$ (see theorem 2.36, 2.28 [299] or theorem 2.38 for weaker conditions). Note

$$\left\| m^\top (\nabla \log p - \nabla \log q) \right\|_2^2 = \|m^\top \nabla \log p\|_2^2 + \|m^\top \nabla \log q\|_2^2 - 2m^\top \nabla \log p \cdot m^\top \nabla \log q$$

thus we have

$$\begin{aligned}
\int_{\mathcal{X}} \left\langle m^\top \nabla \log p, m^\top \nabla \log q \right\rangle \mathrm{d}\mathbb{Q} &= \int_{\mathcal{X}} \left\langle \nabla \log q, mm^\top \nabla \log p \right\rangle \mathrm{d}Q \\
&= \int_{\mathcal{X}} \left\langle \nabla q, mm^\top \nabla \log p \right\rangle \mathrm{d}x \\
&= \int_{\mathcal{X}} \left(\nabla \cdot \left(qmm^\top \nabla \log p \right) - q\nabla \cdot \left(mm^\top \nabla \log p \right) \right) \mathrm{d}x \\
&= -\int_{\mathcal{X}} q\nabla \cdot \left(mm^\top \nabla \log p \right) \mathrm{d}x = -\int_{\mathcal{X}} \nabla \cdot \left(mm^\top \nabla \log p \right) \mathrm{d}\mathbb{Q}.
\end{aligned}$$

6.11.3.2 Proof of 55: DSM as a limit of DKSD

We now consider the the limit in which DKSD converges to DSM:

Theorem 55 (DSM as a limit of DKSD). *Let $\mathbb{Q}$ be a distribution on $\mathbb{R}^d$ with $q > 0$ and suppose $s_p - s_q \in C(\mathbb{R}^d) \cap L^2(\mathbb{Q})$. Let $\Phi_\gamma(s) \equiv \gamma^{-d}\Phi(s/\gamma)$, $\gamma > 0$, $\Phi \in L^1(\mathbb{R}^d)$, $\Phi > 0$ and $\int_{\mathbb{R}^d} \Phi(s)\mathrm{d}s = 1$. Consider the reproducing kernel $k^q_\gamma(x,y) = k_\gamma(x,y)/\sqrt{q(x)q(y)} =$*

$\Phi_\gamma(x-y)/\sqrt{q(x)q(y)}$, *and set* $K^q_\gamma \equiv Bk^q_\gamma$. *Then,*

$$\mathrm{DKSD}_{K^q_\gamma,m}(\mathbb{Q}\|\mathbb{P})^2 \to \mathrm{DSM}_m(\mathbb{Q}\|\mathbb{P}) \quad \text{as } \gamma \to 0.$$

We use the following lemma as a stepping stone.

Lemma 56. *Suppose* $\Phi \in L^1(\mathbb{R}^d)$, $\Phi > 0$ *and* $\int \Phi(s)\,\mathrm{d}s = 1$. *Let* $f, g \in C(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$, *then defining* $K_\gamma \equiv B\Phi_\gamma$ *where* $\Phi_\gamma(s) \equiv \gamma^{-d}\Phi(s/\gamma)$ *and* $\gamma > 0$, *we have*

$$\int\int f(x)^\top K_\gamma(x,y)g(y)\mathrm{d}x\,\mathrm{d}y \to \int f(x)^\top Bg(x)\,\mathrm{d}x, \quad \text{as} \quad \gamma \to 0.$$

Proof. We rewrite

$$\int_{\mathcal{X}}\int_{\mathcal{X}} f(x)^\top B\Phi_\gamma(x-y)g(y)\,\mathrm{d}x\,\mathrm{d}y = \int_{\mathcal{X}}\int_{\mathcal{X}} f(x)^\top Bg(x-s)\mathrm{d}x\,\Phi_\gamma(s)\,\mathrm{d}s = \int_{\mathcal{X}} H(s)\Phi_\gamma(s)\,\mathrm{d}s,$$

where $H : \mathcal{X} \to \mathbb{R}$ is defined by

$$H(s) \equiv \int_{\mathcal{X}} f(x)^\top Bg(x-s)\,\mathrm{d}x = \int_{\mathcal{X}} \langle f(x), Bg(x-s)\rangle_{\mathbb{R}^d}\mathrm{d}x \equiv \int_{\mathcal{X}} \langle f(x), g(x-s)\rangle_B \mathrm{d}x.$$

Since $f, g \in C(\mathbb{R}^d)\cap L^2(\mathbb{R}^d)$, the function $H(s)$ is continuous, bounded, $|H(s)| \le A\|f\|_{L^2(\mathbb{R}^d)}\|g\|_{L^2(\mathbb{R}^d)}$ for a constant $A > 0$ depending only on B, and $H(0) = \int f(x)^\top Bg(x)\,\mathrm{d}x$. Given $\delta > 0$, we can split the integral as follows:

$$\int_{|s|<\delta} H(s)\Phi_\gamma(s)\,\mathrm{d}s + \int_{|s|>\delta} H(s)\Phi_\gamma(s)\,\mathrm{d}s \equiv I_1 + I_2.$$

By continuity, given $\epsilon \in (0,1)$ there exists $\delta > 0$ such that $|H(s) - H(0)| < \epsilon$ for all $|s| < \delta$. Let $I_{<\delta} \equiv \int_{|y|<\delta} \Phi_\gamma(y)\,\mathrm{d}y > 0$ since $\Phi > 0$. Consider

$$\begin{aligned} I_1 - H(0) &= \int_{|s|<\delta} \Phi_\gamma(s)H(s)\mathrm{d}s - H(0) = \int_{|s|<\delta} \Phi_\gamma(s)\left(H(s) - \frac{H(0)}{I_{<\delta}}\right)\mathrm{d}s \\ &= \int_{|s|<\delta} \frac{\Phi_\gamma(s)}{I_{<\delta}}\left(H(s)I_{<\delta} - H(0)\right)\mathrm{d}s. \end{aligned}$$

Clearly $\int \Phi_\gamma(s)\mathrm{d}s = \int \gamma^{-d}\Phi(s/\gamma)\mathrm{d}s = \int \Phi(z)\mathrm{d}z = 1$, since $z \equiv s/\gamma$ implies $\mathrm{d}z = \gamma^{-d}\mathrm{d}s$, so

$$I_{<\delta} = 1 - I_{>\delta} = 1 - \int_{|y|>\delta/\gamma} \Phi(y)\,\mathrm{d}y.$$

Then since Φ is integrable, there exists $\gamma_0(\delta) > 0$ s.t. for $\gamma < \gamma_0(\delta)$ we have $\int_{|y|>\delta/\gamma} \Phi(y)\,\mathrm{d}y < \epsilon$

and thus $0 < 1 - \epsilon < I_{<\delta} < 1$. Therefore, for $\gamma < \gamma_0(\delta)$:

$$\begin{aligned}
|I_1 - H(0)| &= \left| \int_{|s|<\delta} \frac{\Phi_\gamma(s)}{I_{<\delta}} \left(H(s) I_{<\delta} - H(0)\right) \mathrm{d}s \right| \\
&\leq \int_{|s|<\delta} \frac{\Phi_\gamma(s)}{I_{<\delta}} \left| \left(\left(H(s) - H(0)\right) I_{<\delta} + H(0) \left(I_{<\delta} - 1\right)\right) \right| \mathrm{d}s \\
&\leq \int_{|s|<\delta} \frac{\Phi_\gamma(s)}{I_{<\delta}} \left(\left|H(s) - H(0)\right| I_{<\delta} + \left|1 - I_{<\delta}\right| H(0) \right) \mathrm{d}s \\
&\leq \int_{|s|<\delta} \frac{\Phi_\gamma(s)}{I_{<\delta}} \left(\epsilon I_{<\delta} + \epsilon H(0) \right) \mathrm{d}s \\
&\leq \epsilon \int_{|z|<\delta/\gamma} \Phi(z) \mathrm{d}z + H(0)\epsilon \leq \left(1 + H(0)\right) \epsilon.
\end{aligned}$$

For the second term, since H is bounded we have

$$I_2 = \int_{|s|>\delta} H(s)\Phi_\gamma(s)\mathrm{d}s = \int_{|s|>\delta/\gamma} H(\gamma s)\Phi(s)\mathrm{d}s \leq \|H\|_\infty \int_{|s|>\delta/\gamma} \Phi(s)\mathrm{d}s,$$

so that, $|I_2| \leq \|H\|_\infty \epsilon$, for $\gamma < \gamma_0(\delta)$. It follows that

$$\begin{aligned}
\left| \int \int f(x)^\top K_\gamma(x,y) g(y) \mathrm{d}x \, \mathrm{d}y - \int f(x)^\top B g(x) \, \mathrm{d}x \right| &= \left| \int H(s)\Phi_\gamma(s)\mathrm{d}s - H(0) \right| \\
&= |I_1 + I_2 - H(0)| \\
&\leq |I_1 - H(0)| + |I_2| \to 0,
\end{aligned}$$

as $\gamma \to 0$ as required. ■

We note that $f \in L^2(\mathbb{Q})$ if and only if $f\sqrt{q} \in L^2(\mathbb{R}^d)$. Therefore applying the previous result, we have that

$$\begin{aligned}
\int_{\mathcal{X}} \int_{\mathcal{X}} f(x)^\top K_\gamma^q(x,y) g(y) \, \mathrm{d}\mathbb{Q}(x) \, \mathrm{d}\mathbb{Q}(y) &= \int_{\mathcal{X}} \int_{\mathcal{X}} \left(\sqrt{q(x)} f(x) \right)^\top K_\gamma(x,y) \left(g(y) \sqrt{q(y)} \right) \mathrm{d}x \mathrm{d}y \\
&\to \int_{\mathcal{X}} f(x)^\top B g(x) \mathrm{d}\mathbb{Q}(x), \quad \text{as} \quad \gamma \to 0.
\end{aligned}$$

Note that if k is a (scalar) kernel function, then $(x,y) \mapsto r(x)k(x,y)r(y)$ is a kernel for any function $r : \mathcal{X} \to \mathbb{R}$, and thus k_γ^q defines a sequence of kernels parametrised by a scale parameter $\gamma > 0$. It follows that the sequence of DKSD paramaterised by K_γ^q

$$\mathrm{DKSD}_{K_\gamma^q, m}(\mathbb{Q} \| \mathbb{P})^2 = \int_{\mathcal{X}} \int_{\mathcal{X}} q(x) \delta_{p,q}(x)^\top K_\gamma^q(x,y) \delta_{p,q}(y) q(y) \mathrm{d}x \mathrm{d}y$$

converges to DSM with inner product $\langle \cdot, \cdot \rangle_B \equiv \langle \cdot, B \cdot \rangle_2$ on $\mathbb{R}^d$.

$$\mathrm{DSM}_m(\mathbb{Q} \| \mathbb{P}) = \int_{\mathcal{X}} \delta_{q,p}(x)^\top B \delta_{q,p}(x) \mathrm{d}\mathbb{Q} = \int_{\mathcal{X}} \| m^\top \left(\nabla \log p - \nabla \log q \right) \|_B^2 \mathrm{d}\mathbb{Q}$$

6.12 Information Semi-Metrics of Minimum Stein Discrepancy Estimators

In this section, we derive expressions for the metric tensor of DKSD and DSM (recall §5.1.3).

6.12.1 Proof of 39: Information Semi-Metric of Diffusion Kernel Stein Discrepancy

From 36 we have

$$\mathrm{DKSD}_{K,m}(\mathbb{P}_\alpha, \mathbb{P}_\theta)^2 = \int_{\mathcal{X}} \int_{\mathcal{X}} p_\alpha(x) \delta_{p_\theta, p_\alpha}(x)^\top K(x,y) \delta_{p_\theta, p_\alpha}(y) p_\alpha(y) \mathrm{d}x \mathrm{d}y$$

where $\delta_{p_\theta, p_\alpha} = m_\theta^\top \left(\nabla \log p_\theta - \nabla \log p_\alpha \right)$. Thus

$$\begin{aligned}
\partial_{\alpha^i} \partial_{\theta^j} \mathrm{DKSD}_{K,m}(\mathbb{P}_\alpha, \mathbb{P}_\theta)^2 &= \partial_{\alpha^i} \partial_{\theta^j} \int_{\mathcal{X}} \int_{\mathcal{X}} p_\alpha(x) \delta_{p_\theta, p_\alpha}(x)^\top K(x,y) \delta_{p_\theta, p_\alpha}(y) p_\alpha(y) \mathrm{d}x \mathrm{d}y \\
&= \partial_{\alpha^i} \int_{\mathcal{X}} \int_{\mathcal{X}} p_\alpha(x) \partial_{\theta^j} \delta_{p_\theta, p_\alpha}(x)^\top K(x,y) \delta_{p_\theta, p_\alpha}(y) p_\alpha(y) \mathrm{d}x \mathrm{d}y \\
&+ \partial_{\alpha^i} \int_{\mathcal{X}} \int_{\mathcal{X}} p_\alpha(x) \delta_{p_\theta, p_\alpha}(x)^\top K(x,y) \partial_{\theta^j} \delta_{p_\theta, p_\alpha}(y) p_\alpha(y) \mathrm{d}x \mathrm{d}y,
\end{aligned}$$

and using $\delta_{p_\theta, p_\theta} = 0$, we get:

$$\begin{aligned}
&\partial_{\alpha^i} \int_{\mathcal{X}} \int_{\mathcal{X}} p_\alpha(x) \partial_{\theta^j} \delta_{p_\theta, p_\alpha}(x)^\top K(x,y) \delta_{p_\theta, p_\alpha}(y) p_\alpha(y) \mathrm{d}x \mathrm{d}y \Big|_{\alpha=\theta} \\
&= \partial_{\alpha^i} \int_{\mathcal{X}} \int_{\mathcal{X}} p_\alpha(x) \left(\partial_{\theta^j} m_\theta^\top (\nabla \log p_\theta - \nabla \log p_\alpha) + m_\theta^\top \partial_{\theta^j} \nabla \log p_\theta \right)^\top K(x,y) \delta_{p_\theta, p_\alpha}(y) p_\alpha(y) \mathrm{d}x \mathrm{d}y \Big|_{\alpha=\theta} \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} p_\alpha(x) \left(m_\theta^\top \partial_{\theta^j} \nabla \log p_\theta \right)^\top K(x,y) \partial_{\alpha^i} \delta_{p_\theta, p_\alpha}(y) p_\alpha(y) \mathrm{d}x \mathrm{d}y \Big|_{\alpha=\theta} \\
&= - \int_{\mathcal{X}} \int_{\mathcal{X}} p_\alpha(x) \left(m_\theta^\top \partial_{\theta^j} \nabla \log p_\theta \right)^\top K(x,y) \left(m_\theta^\top \partial_{\alpha^i} \nabla \log p_\alpha \right)(y) p_\alpha(y) \mathrm{d}x \mathrm{d}y \Big|_{\alpha=\theta} \\
&= - \int_{\mathcal{X}} \int_{\mathcal{X}} \left(m_\theta^\top \partial_{\theta^j} \nabla \log p_\theta \right)^\top (x) K(x,y) \left(m_\theta^\top \partial_{\theta^i} \nabla \log p_\theta \right)(y) \mathrm{d}\mathbb{P}_\theta(x) \mathrm{d}\mathbb{P}_\theta(y).
\end{aligned}$$

Similarly, we also get:

$$\begin{aligned}
&\partial_{\alpha^i} \int_{\mathcal{X}} \int_{\mathcal{X}} p_\alpha(x) \delta_{p_\theta, p_\alpha}(x)^\top K(x,y) \partial_{\theta^j} \delta_{p_\theta, p_\alpha}(y) p_\alpha(y) \mathrm{d}x \mathrm{d}y \big|_{\alpha=\theta} \\
&= - \int_{\mathcal{X}} \int_{\mathcal{X}} \left(m_\theta^\top \partial_{\theta^i} \nabla \log p_\theta \right)^\top (x) K(x,y) \left(m_\theta^\top \partial_{\theta^j} \nabla \log p_\theta \right)(y) \mathrm{d}\mathbb{P}_\theta(x) \mathrm{d}\mathbb{P}_\theta(y) \\
&= - \int_{\mathcal{X}} \int_{\mathcal{X}} \left(m_\theta^\top \partial_{\theta^i} \nabla \log p_\theta \right)^\top (y) K(y,x) \left(m_\theta^\top \partial_{\theta^j} \nabla \log p_\theta \right)(x) \mathrm{d}\mathbb{P}_\theta(y) \mathrm{d}\mathbb{P}_\theta(x) \\
&= - \int_{\mathcal{X}} \int_{\mathcal{X}} \left(m_\theta^\top \partial_{\theta^i} \nabla \log p_\theta \right)^\top (y) K(x,y)^\top \left(m_\theta^\top \partial_{\theta^j} \nabla \log p_\theta \right)(x) \mathrm{d}\mathbb{P}_\theta(y) \mathrm{d}\mathbb{P}_\theta(x) \\
&= - \int_{\mathcal{X}} \int_{\mathcal{X}} \left(m_\theta^\top \partial_{\theta^j} \nabla \log p_\theta \right)(x)^\top K(x,y) \left(m_\theta^\top \partial_{\theta^i} \nabla \log p_\theta \right)(y) \mathrm{d}\mathbb{P}_\theta(y) \mathrm{d}\mathbb{P}_\theta(x).
\end{aligned}$$

Hence, we conclude that

$$\tfrac{1}{2} \partial_{\alpha^i} \partial_{\theta^j} \operatorname{DKSD}_{K,m}(\mathbb{P}_\alpha, \mathbb{P}_\theta)^2 = - \int_{\mathcal{X}} \int_{\mathcal{X}} \left(m_\theta^\top \partial_{\theta^j} \nabla \log p_\theta \right)(x)^\top K(x,y) \left(m_\theta^\top \partial_{\theta^i} \nabla \log p_\theta \right)(y) \mathrm{d}\mathbb{P}_\theta(y) \mathrm{d}\mathbb{P}_\theta(x)$$

The information tensor is positive semi-definite. Indeed writing $V_\theta(y) \equiv m_\theta^\top(y) \nabla_y \langle v, \nabla_\theta \log p_\theta \rangle$:

$$\begin{aligned}
\langle v, g(\theta) v \rangle &= v^i g_{ij}(\theta) v^j \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} \left(m_\theta^\top(x) \nabla_x \langle v, \nabla_\theta \log p_\theta \rangle \right)^\top K(x,y) \left(m_\theta^\top(y) \nabla_y \langle v, \nabla_\theta \log p_\theta \rangle \right) \mathrm{d}\mathbb{P}_\theta(x) \mathrm{d}\mathbb{P}_\theta(y) \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} \left\langle m_\theta^\top(x) \nabla_x \langle v, \nabla_\theta \log p_\theta \rangle, K(x,y) m_\theta^\top(y) \nabla_y \langle v, \nabla_\theta \log p_\theta \rangle \right\rangle \mathrm{d}\mathbb{P}_\theta(x) \mathrm{d}\mathbb{P}_\theta(y) \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} \langle V_\theta(x), K(x,y) V_\theta(y) \rangle \, \mathrm{d}\mathbb{P}_\theta(x) \mathrm{d}\mathbb{P}_\theta(y) \geq 0
\end{aligned}$$

since K is IPD.

6.12.2 Proof of 40: Information Semi-Metric of Diffusion Score Matching

Proof. The information metric is given by $g(\theta)_{ij} = -\frac{1}{2} \frac{\partial^2}{\partial \alpha^i \partial \theta^j} \operatorname{DSM}(p_\alpha \| p_\theta)|_{\alpha=\theta}$. Recall

$$\operatorname{DSM}(p_\alpha \| p_\theta) = \int_{\mathcal{X}} \left\| m^\top \left(\nabla \log p_\theta - \nabla \log p_\alpha \right) \right\|_2^2 p_\alpha \mathrm{d}x.$$

Moreover

$$\tfrac{1}{2} \partial_{\alpha^i} \partial_{\theta^j} \operatorname{DSM}(p_\alpha \| p_\theta) \big|_{\alpha=\theta} = \tfrac{1}{2} \partial_{\alpha^i} \partial_{\theta^j} \int_{\mathcal{X}} \left\| m^\top \left(\nabla \log p_\theta - \nabla \log p_\alpha \right) \right\|_2^2 p_\alpha \mathrm{d}x \big|_{\alpha=\theta}$$

$$
\begin{aligned}
&= \partial_{\alpha^i} \int_{\mathcal{X}} \left(m^\top \left(\nabla \log p_\theta - \nabla \log p_\alpha\right)\right) \cdot \left(m^\top \partial_{\theta^j} \nabla \log p_\theta\right) p_\alpha \mathrm{d}x \big|_{\alpha=\theta} \\
&= \int_{\mathcal{X}} \left(m^\top \left(\nabla \log p_\theta - \nabla \log p_\alpha\right)\right) \cdot \left(m^\top \partial_{\theta^j} \nabla \log p_\theta\right) \partial_{\alpha^i} p_\alpha \mathrm{d}x \big|_{\alpha=\theta} \\
&- \int_{\mathcal{X}} \left(m^\top \partial_{\alpha^i} \nabla \log p_\alpha\right) \cdot \left(m^\top \partial_{\theta^j} \nabla \log p_\theta\right) p_\alpha \mathrm{d}x \big|_{\alpha=\theta} \\
&= -\int_{\mathcal{X}} \left(m^\top \partial_{\theta^i} \nabla \log p_\theta\right) \cdot \left(m^\top \partial_{\theta^j} \nabla \log p_\theta\right) \mathrm{d}\mathbb{P}_\theta.
\end{aligned}
$$

Finally g is semi-positive definite,

$$
\begin{aligned}
\langle v, g(\theta) v \rangle = v^i g_{ij}(\theta) v^j &= \int_{\mathcal{X}} v^i m_{rs}^\top \partial_{x^s} \partial_{\theta^i} \log p_\theta m_{rl}^\top \partial_{x^l} \partial_{\theta^j} \log p_\theta v^j \mathrm{d}\mathbb{P}_\theta \\
&= \int_{\mathcal{X}} m_{rs}^\top \partial_{x^s} \langle v, \nabla_\theta \log p_\theta \rangle \, m_{rl}^\top \partial_{x^l} \langle v, \nabla_\theta \log p_\theta \rangle \, \mathrm{d}\mathbb{P}_\theta \\
&= \int_{\mathcal{X}} \left\langle m^\top \nabla_x \langle v, \nabla_\theta \log p_\theta \rangle, m^\top \nabla_x \langle v, \nabla_\theta \log p_\theta \rangle \right\rangle \mathrm{d}\mathbb{P}_\theta \\
&= \int_{\mathcal{X}} \| m^\top \nabla_x \langle v, \nabla_\theta \log p_\theta \rangle \|^2 \mathrm{d}\mathbb{P}_\theta \geq 0
\end{aligned}
$$

■

6.13 Proofs of Consistency and Asymptotic Normality for minimum Stein Discrepancy Estimators

In this section we prove several results concerning the consistency and asymptotic normality of DKSD and DSM estimators.

6.13.1 Diffusion Kernel Stein Discrepancies

Given the Stein kernel (6.6) we want to estimate $\theta_*^{\mathrm{DKSD}} \equiv \operatorname{argmin}_{\theta \in \Theta} \mathrm{DKSD}_{K,m}(\mathbb{Q}, \mathbb{P}_\theta)^2 = \operatorname{argmin}_{\theta \in \Theta} \int k_\theta^0(x, y) \mathbb{Q}(\mathrm{d}x) \mathbb{Q}(\mathrm{d}y)$ using a sequence of estimators $\hat{\theta}_n^{\mathrm{DKSD}} \in \operatorname{argmin}_{\theta \in \Theta} \widehat{\mathrm{DKSD}}_{K,m}(\mathbb{Q}, \mathbb{P}_\theta)^2$ that minimise the U-statistic approximation (6.7). We will assume we are in the specified setting $\mathbb{Q} = \mathbb{P}_{\theta_*^{\mathrm{DKSD}}} \in \mathcal{P}_\Theta$. In the misspecified setting it is necessary to further assume the existence of a unique minimiser.

6.13.1.1 Strong Consistency

We first prove a general strong consistency result based on an equicontinuity assumption:

Lemma 57. *Let $\mathcal{X} = \mathbb{R}^d$. Suppose $\{\theta \mapsto k^0_\theta(x,y)\}, \{\theta \mapsto \mathbb{Q}_z k^0_\theta(x,z)\}$ are equicontinuous on any compact subset $C \subset \Theta$ for x, y in a sequence of sets whose union has full $\mathbb{Q}$-measure, and $\|s_{p_\theta}(x)\| \leq f_1(x)$, $\|\nabla_x \cdot m_\theta(x)\| \leq f_2(x)$, $\|\nabla_x \cdot (m_\theta(x)K(x,y))\| \leq f_3(x,y)$, $|\operatorname{tr}[m(y)\nabla_y\nabla_x \cdot (m(x)K)]| \leq f_4(x,y)$ hold on C, where $f_1(x)\sqrt{K(x,x)_{ii}} \in L^1(\mathbb{Q})$, and $f_4, f_3 f_2, f_1 f_3 \in L^1(\mathbb{Q}\otimes\mathbb{Q})$. Assume further that $\theta \mapsto \mathbb{P}_\theta$ is injective. Then we have a unique minimiser θ^{DKSD}_*, and if either Θ is compact, or $\theta^{\mathrm{DKSD}}_* \in \mathit{int}(\Theta)$ and Θ and $\theta \mapsto \widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_\theta)^2$ are convex, then $\hat{\theta}^{\mathrm{DKSD}}_n$ is strongly consistent.*

Proof. Note $\mathrm{DKSD}_{K,m}(\mathbb{Q}, \mathbb{P}_\theta)^2 = 0$ iff $\mathbb{P}_\theta = \mathbb{P}_{\theta^{\mathrm{DKSD}}_*}$ by 36, which implies $\theta = \theta^{\mathrm{DKSD}}_*$ since $\theta \mapsto \mathbb{P}_\theta$ is injective. Thus we have a unique minimiser at θ^{DKSD}_*.

Suppose first Θ is compact and take $C = \Theta$. Note

$$\begin{aligned}
|k^0(x,y)| \leq & |\langle s_p(x), K s_p(y)\rangle| + |\langle \nabla_y \cdot (m(y)K), s_p(x)\rangle| + |\langle \nabla_x \cdot (m(x)K), s_p(y)\rangle| \\
& + |\langle \nabla_x \cdot (m(x)K), \nabla_y \cdot m\rangle| + |\operatorname{tr}[m(y)\nabla_y\nabla_x \cdot (m(x)K)]| \\
\leq & |\langle s_p(x), K s_p(y)\rangle| + f_3(y,x)f_1(x) + f_3(x,y)f_1(y) + f_3(x,y)f_2(y) + f_4(x,y),
\end{aligned}$$

From the reproducing property $f(x) = \langle f, K(\cdot,x)v\rangle_{\mathcal{H}^d}$, for any $f \in \mathcal{H}^d$, $v \in \mathbb{R}^d$. Using $K(y,x) = K(x,y)^\top$ we have $K(\cdot,x)_{,i} = K(x,\cdot)_{i,}$, where $K(\cdot,x)_{,i}$ and $K(x,\cdot)_{i,}$ denote the i^{th} column and row respectively, which implies that $K(x,\cdot)_{i,}, K(\cdot,x)_{,i} \in \mathcal{H}^d$ and $f(x)_i = \langle f, K(\cdot,x)_{,i}\rangle_{\mathcal{H}^d}$. Choosing $f = K(\cdot,y)_{,j}$ implies

$$\begin{aligned}
K(x,y)_{ij} = \langle K(\cdot,y)_{,j}, K(\cdot,x)_{,i}\rangle_{\mathcal{H}^d} &\leq \|K(\cdot,y)_{,j}\|_{\mathcal{H}^d}\|K(\cdot,x)_{,i}\|_{\mathcal{H}^d} \\
&= \sqrt{\langle K(\cdot,y)_{,j}, K(\cdot,y)_{,j}\rangle_{\mathcal{H}^d}}\sqrt{\langle K(\cdot,x)_{,i}, K(\cdot,x)_{,i}\rangle_{\mathcal{H}^d}} \\
&= \sqrt{K(y,y)_{jj}}\sqrt{K(x,x)_{ii}}.
\end{aligned}$$

It follows that

$$\begin{aligned}
\langle s_p(x), K s_p(y)\rangle = (s_p)_i(x)K(x,y)_{ij}(s_p)_j(y) &\leq (s_p)_i(x)\sqrt{K(x,x)_{ii}}\sqrt{K(y,y)_{jj}}(s_p)_j(y) \\
&\leq \|s_p(x)\|_\infty\sqrt{K(x,x)_{ii}}\sqrt{K(y,y)_{jj}}\|s_p(y)\|_\infty \\
&\leq C f_1(x)\sqrt{K(x,x)_{ii}}\sqrt{K(y,y)_{jj}}f_1(y),
\end{aligned}$$

where the constant $C > 0$ arises from the norm-equivalence of $\|s_p(y)\|$ and $\|s_p(y)\|_\infty$. Hence k^0

is integrable. Thus by theorem 1 [364],

$$\sup_{\theta} \left| \widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_\theta)^2 - \mathrm{DKSD}_{K,m}(\mathbb{Q}, \mathbb{P}_\theta)^2 \right| \xrightarrow{a.s.} 0$$

and $\theta \mapsto \mathrm{DKSD}_{K,m}(\mathbb{Q}, \mathbb{P}_\theta)^2$ are continuous. By theorem 2.1 [277] then $\hat{\theta}_n^{\mathrm{DKSD}} \xrightarrow{a.s.} \theta_*^{\mathrm{DKSD}}$.

On the other hand, if Θ is convex we follow a similar strategy to the proof of theorem 2.7 [277]. Since $\theta_*^{\mathrm{DKSD}} \in \mathrm{int}(\Theta)$, we can find a $\epsilon > 0$ for which $C = \overline{B}(\theta_*^{\mathrm{DKSD}}, 2\epsilon) \subset \Theta$ is a closed ball containing θ_*^{DKSD} (which is compact since $\Theta \subset \mathbb{R}^m$). Using the compact case, we know any sequence of estimators $\tilde{\theta}_n^{\mathrm{DKSD}} \in \mathrm{argmin}_{\theta \in C} \widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_\theta)^2$ is strongly consistent for θ_*^{DKSD}. In particular, there exists N_0 a.s. s.t. for $n > N_0$, $\|\tilde{\theta}_n^{\mathrm{DKSD}} - \theta_*^{\mathrm{DKSD}}\| < \epsilon$. If $\theta \notin C$, there exists $\lambda \in [0, 1)$ s.t. $\lambda\tilde{\theta}_n^{\mathrm{DKSD}} + (1-\lambda)\theta$ lies on the boundary of the closed ball C. Using convexity and the fact $\tilde{\theta}_n^{\mathrm{DKSD}}$ is a minimiser over C,

$$\begin{aligned}
&\widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_{\tilde{\theta}_n^{\mathrm{DKSD}}})^2 \\
&\quad \leq \widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_{\lambda\tilde{\theta}_n^{\mathrm{DKSD}} + (1-\lambda)\theta})^2 \\
&\quad \leq \lambda\widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_{\tilde{\theta}_n^{\mathrm{DKSD}}})^2 + (1-\lambda)\widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_\theta)^2
\end{aligned}$$

which implies $\widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_{\tilde{\theta}_n^{\mathrm{DKSD}}})^2 \leq \widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_\theta)^2$ and $\tilde{\theta}_n^{\mathrm{DKSD}}$ is the global minimum of $\theta \mapsto \widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_\theta)^2$ for $n > N_0$. ■

When k^0 is Fréchet differentiable on Θ equicontinuity can be obtained using the Mean value theorem, which simplifies the assumptions under which strong consistency holds.

We now prove our main result for consistency of minimum DKSD estimators: 41:

Proof. Let $\|K\| + \|\nabla_x K\| + \|\nabla_x \nabla_y K\| \leq K_\infty$. Note $\|\nabla_y \cdot (m(y)K)\| \leq 2f_2(y)K_\infty$ and $|\operatorname{tr}[m(y)\nabla_y \nabla_x \cdot (m(x)K)]| \leq 2f_2(y)f_2(x)K_\infty$ so

$$|k_\theta^0(x, y)| \leq f_1(x)K_\infty f_1(y) + 2f_2(x)K_\infty f_1(y) + 2f_2(y)K_\infty f_1(x) + 3K_\infty f_2(x)f_2(y)$$

which is symmetric and integrable by assumption. Let S_m, $m = 1, 2, \ldots$ be an increasing sequence of closed balls in $\mathbb{R}^d$, such that $\cup_{m=1}^\infty S_m = \mathbb{R}^d$. Moreover,

$$\begin{aligned}
\|\nabla_\theta \langle s_p(x), Ks_p(y)\rangle\| &\leq g_1(x)f_1(y)K_\infty + g_1(y)f_1(x)K_\infty \\
\|\nabla_\theta \langle \nabla_y \cdot (m(y)K), s_p(x)\rangle\| &\leq 2K_\infty g_2(y)f_1(x) + 2f_2(y)g_1(x)K_\infty \\
\|\nabla_\theta \langle \nabla_x \cdot (m(x)K), \nabla_y \cdot m\rangle\| &\leq 2K_\infty g_2(x)f_2(y) + 2K_\infty f_2(x)g_2(y) \\
\|\nabla_\theta \operatorname{tr}[m(y)\nabla_y \nabla_x \cdot (m(x)K)]\| &\leq 2K_\infty g_2(y)f_2(x) + 2K_\infty f_2(y)g_2(x)
\end{aligned}$$

thus $\|\nabla_\theta k^0_\theta(x,y)\|$ is bounded above by a continuous integrable symmetric function, $(x,y) \mapsto s(x,y)$, which attains a maximum on the compact spaces $S_m \times S_m$. By the MVT applied on the $\mathbb{R}^m$-open neighbourhood of Θ, $|k^0_\theta(x,y) - k^0_\alpha(x,y)| \leq \|\nabla_\theta k^0_\theta(x,y)\|\|\theta - \alpha\| \leq s(x,y)\|\theta - \alpha\| \leq \max_{x,y\in S_m} s(x,y)\|\theta - \alpha\|$, and $k^0_\theta(x,y)$ is equicontinuous in $\theta \in C$ for $x, y \in S_m$. Similarly, since s is integrable, $|\int_{\mathcal{X}} k^0_\theta(x,y)\mathbb{Q}(dy) - \int_{\mathcal{X}} k^0_\alpha(x,z)\mathbb{Q}(dz)| \leq \|\nabla_\theta \int_{\mathcal{X}} k^0_\theta(x,z)\mathrm{d}\mathbb{Q}(z)\|\|\theta - \alpha\| \leq \int_{\mathcal{X}} \|\nabla_\theta k^0_\theta(x,z)\|\mathrm{d}\mathbb{Q}(z)\|\theta - \alpha\| \leq \max_{x\in S_m} \mathbb{Q}_z s(x,z)\|\theta - \alpha\| \leq$ is equicontinuous in $\theta \in C$ for $x \in S_m$. The rest follows as in the previous proposition. ■

6.13.1.2 Proof of 42: Asymptotic Normality

Proof. Note that $\nabla_\theta \widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_\theta)^2 = \frac{1}{N(N-1)} \sum_{i\neq j} \nabla_\theta k^0_\theta(X_i, X_j)$. Let $\mu(\theta) \equiv \mathbb{Q} \otimes \mathbb{Q}[\nabla_\theta k^0_\theta]$. Assumptions 1 and 2 imply that $\mathbb{Q} \otimes \mathbb{Q}[\|\nabla_\theta k^0_\theta\|^2] < \infty$. By [155, Theorem 7.1] it follows that

$$\sqrt{n}\left(\nabla_\theta \widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_\theta)^2 - \mu(\theta)\right) \xrightarrow{d} \mathcal{N}(0, 4\Sigma(\theta))$$

where

$$\begin{aligned}\Sigma &= \mathbb{Q}\left[\mathbb{Q}_2\left[\nabla_\theta k^0_\theta - \mu(\theta)\right] \otimes \mathbb{Q}_2\left[\nabla_\theta k^0_\theta - \mu(\theta)\right]\right] \\ &= \int_{\mathcal{X}} \left(\int_{\mathcal{X}} \nabla_\theta k^0_\theta(x,y)\mathrm{d}\mathbb{Q}(y) - \mu(\theta)\right) \otimes \left(\int_{\mathcal{X}} \nabla_\theta k^0_\theta(x,z)\mathrm{d}\mathbb{Q}(z) - \mu(\theta)\right) \mathrm{d}\mathbb{Q}(x)\end{aligned}$$

Note that $\mu(\theta^{\mathrm{DKSD}}_*) = \mathbb{Q} \otimes \mathbb{Q}[\nabla_\theta k^0_\theta|_{\theta^{\mathrm{DKSD}}_*}] = \nabla_\theta \left(\mathbb{Q} \otimes \mathbb{Q}[k^0_\theta]\right)|_{\theta=\theta^{\mathrm{DKSD}}_*}$, and if $\mathbb{Q} \otimes \mathbb{Q}[k^0_\theta]$ is differentiable around θ^{DKSD}_*, then the first order optimality condition implies $\mu(\theta^{\mathrm{DKSD}}_*) = 0$.

Consider now $\nabla_\theta \nabla_\theta \widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}, \mathbb{P}_\theta)^2 = \frac{1}{n(n-1)} \sum_{i\neq j} \nabla_\theta \nabla_\theta k^0_\theta(X_i, X_j)$. Note

$$\begin{aligned}\|\nabla_\theta \nabla_\theta \nabla_\theta \langle s_p(x), K s_p(y)\rangle\| &\lesssim g_1(x) K_\infty f_1(y) + f_1(x) K_\infty g_1(y) \\ \|\nabla_\theta \nabla_\theta \nabla_\theta \langle \nabla_y \cdot (m(y)K), s_p(x)\rangle\| &\lesssim g_2(y) K_\infty f_1(x) + f_2(y) K_\infty g_1(x) \\ \|\nabla_\theta \nabla_\theta \nabla_\theta \langle \nabla_x \cdot (m(x)K), \nabla_y \cdot m\rangle\| &\lesssim f_2(y) K_\infty g_2(x) + g_2(y) K_\infty f_2(x) \\ \|\nabla_\theta \nabla_\theta \nabla_\theta \operatorname{tr}[m(y)\nabla_y \nabla_x \cdot (m(x)K)]\| &\lesssim g_2(y) K_\infty f_2(x) + f_2(y) K_\infty g_2(x)\end{aligned}$$

Hence by Assumptions 1-4 $\|\nabla_\theta \nabla_\theta \nabla_\theta k^0_\theta\|$ is bounded above by a continuous integrable symmetric function and we can apply the MVT to show equicontinuity as in the proof above. Moreover the conditions of [364, Theorem 1] hold for the components of $\nabla_\theta \nabla_\theta k^0_\theta$, so that $\sup_{\theta\in\mathcal{N}} \left|\frac{1}{n(n-1)} \sum_{i\neq j} \partial_{\theta^a}\partial_{\theta^b} k^0_\theta(X_i, X_j) - \mathbb{Q} \otimes \mathbb{Q}\partial_{\theta^a}\partial_{\theta^b} k^0_\theta\right| \xrightarrow{a.s.} 0$ as $n \to \infty$, for all a and b.

Finally we observe that $\mathbb{Q} \otimes \mathbb{Q}\partial_{\theta^a}\partial_{\theta^b} k^0_\theta\big|_{\theta=\theta^{\mathrm{DKSD}}_*} = g_{ab}(\theta^{\mathrm{DKSD}}_*)$, where g is the information

metric associated with $\mathrm{DKSD}_{K,m}$. Indeed using $\delta_{p,q} = 0$ if $p = q$

$$
\begin{aligned}
&\mathbb{Q} \otimes \mathbb{Q} \partial_{\theta^a} \partial_{\theta^b} k_\theta^0 \big|_{\theta=\theta_*^{\mathrm{DKSD}}} \\
&= \partial_{\theta^a} \partial_{\theta^b} \int_{\mathcal{X}} \int_{\mathcal{X}} p_{\theta_*^{\mathrm{DKSD}}}(x) \delta_{p_\theta, p_{\theta_*^{\mathrm{DKSD}}}}(x)^\top K(x,y) \delta_{p_\theta, p_{\theta_*^{\mathrm{DKSD}}}}(y) p_{\theta_*^{\mathrm{DKSD}}}(y) \mathrm{d}x\mathrm{d}y \big|_{\theta=\theta_*^{\mathrm{DKSD}}} \\
&= \partial_{\theta^a} \int_{\mathcal{X}} \int_{\mathcal{X}} p_{\theta_*^{\mathrm{DKSD}}}(x) \partial_{\theta^b} \delta_{p_\theta, p_{\theta_*^{\mathrm{DKSD}}}}(x)^\top K(x,y) \delta_{p_\theta, p_{\theta_*^{\mathrm{DKSD}}}}(y) p_{\theta_*^{\mathrm{DKSD}}}(y) \mathrm{d}x\mathrm{d}y \big|_{\theta=\theta_*^{\mathrm{DKSD}}} \\
&+ \partial_{\theta^a} \int_{\mathcal{X}} \int_{\mathcal{X}} p_{\theta_*^{\mathrm{DKSD}}}(x) \delta_{p_\theta, p_{\theta_*^{\mathrm{DKSD}}}}(x)^\top K(x,y) \partial_{\theta^b} \delta_{p_\theta, p_{\theta_*^{\mathrm{DKSD}}}}(y) p_{\theta_*^{\mathrm{DKSD}}}(y) \mathrm{d}x\mathrm{d}y \big|_{\theta=\theta_*^{\mathrm{DKSD}}} \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} p_{\theta_*^{\mathrm{DKSD}}}(x) \partial_{\theta^b} \delta_{p_\theta, p_{\theta_*^{\mathrm{DKSD}}}}(x)^\top K(x,y) \partial_{\theta^a} \delta_{p_\theta, p_{\theta_*^{\mathrm{DKSD}}}}(y) p_{\theta_*^{\mathrm{DKSD}}}(y) \mathrm{d}x\mathrm{d}y \big|_{\theta=\theta_*^{\mathrm{DKSD}}} \\
&+ \int_{\mathcal{X}} \int_{\mathcal{X}} p_{\theta_*^{\mathrm{DKSD}}}(x) \partial_{\theta^a} \delta_{p_\theta, p_{\theta_*^{\mathrm{DKSD}}}}(x)^\top K(x,y) \partial_{\theta^b} \delta_{p_\theta, p_{\theta_*^{\mathrm{DKSD}}}}(y) p_{\theta_*^{\mathrm{DKSD}}}(y) \mathrm{d}x\mathrm{d}y \big|_{\theta=\theta_*^{\mathrm{DKSD}}} \\
&= 2 \int_{\mathcal{X}} \int_{\mathcal{X}} \left(m_{\theta_*^{\mathrm{DKSD}}}^\top(x) \nabla_x \partial_{\theta^j{}_*^{\mathrm{DKSD}}} \log p_{\theta_*^{\mathrm{DKSD}}} \right)^\top K(x,y) \\
&\qquad \left(m_{\theta_*^{\mathrm{DKSD}}}^\top(y) \nabla_y \partial_{\theta^i{}_*^{\mathrm{DKSD}}} \log p_{\theta_*^{\mathrm{DKSD}}} \right) \mathrm{d}\mathbb{P}_{\theta_*^{\mathrm{DKSD}}}(x) \mathrm{d}\mathbb{P}_{\theta_*^{\mathrm{DKSD}}}(y),
\end{aligned}
$$

so $\mathbb{Q} \otimes \mathbb{Q} \partial_{\theta^a} \partial_{\theta^b} k_\theta^0 \big|_{\theta=\theta_*^{\mathrm{DKSD}}} = g_{ab}(\theta_*^{\mathrm{DKSD}})$. The conditions of [277, Theorem 3.1] hold, from which the advertised result follows. ■

6.13.2 Diffusion Score Matching

Recall that the DSM is given by:

$$
\mathrm{DSM}(\mathbb{Q} \| \mathbb{P}_\theta) = \int_{\mathcal{X}} \left(\left\| m^\top \nabla_x \log p_\theta \right\|_2^2 + \| m^\top \nabla \log q \|_2^2 + 2 \nabla \cdot \left(m m^\top \nabla \log p_\theta \right) \right) \mathrm{d}\mathbb{Q}
$$

and we wish to estimate

$$
\begin{aligned}
\theta_*^{\mathrm{DSM}} &= \mathrm{argmin}_{\theta \in \Theta} \int_{\mathcal{X}} \left(\left\| m^\top \nabla_x \log p_\theta \right\|_2^2 + 2 \nabla \cdot \left(m m^\top \nabla \log p_\theta \right) \right) \mathrm{d}\mathbb{Q} && \text{(6.23a)} \\
&\equiv \mathrm{argmin}_{\theta \in \Theta} \int_{\mathcal{X}} F_\theta \mathrm{d}\mathbb{Q} && \text{(6.23b)}
\end{aligned}
$$

with a sequence of M-estimators $\hat{\theta}_n^{\mathrm{DSM}} = \mathrm{argmin}_{\theta \in \Theta} \frac{1}{n} \sum_i^n F_\theta(X_i)$. Recall also we have

$$
F_\theta(x) = \left\| m^\top \nabla_x \log p_\theta \right\|_2^2 + 2 \left\langle \nabla \cdot (m m^\top), \nabla \log p_\theta \right\rangle + 2 \operatorname{tr} \left[m m^\top \nabla^2 \log p_\theta \right].
$$

We will have a unique minimiser θ_*^{DSM} whenever the map $\theta \mapsto \mathbb{P}_\theta$ is injective.

6.13.2.1 Weak Consistency of DSM

Theorem 58 (**Weak Consistency of DSM**). *Suppose $\mathcal{X}$ be open subset of $\mathbb{R}^d$, and $\Theta \subset \mathbb{R}^m$. Suppose $\log p_\theta(\cdot)$ is $C^2(\mathcal{X})$ and $m \in C^1(\mathcal{X})$, and $\|\nabla_x \log p_\theta(x)\| \leq f_1(x)$. Suppose also that $\|\nabla_x \nabla_x \log p_\theta(x)| \leq f_2(x)$ on any compact set $C \subset \Theta$, where $\|m^\top\| f_1 \in L^2(\mathbb{Q})$, $\|\nabla \cdot (mm^\top)\| f_1 \in L^1(\mathbb{Q})$, $\|mm^\top\|_\infty f_2 \in L^1(\mathbb{Q})$. If either Θ is compact, or Θ and $\theta \mapsto F_\theta$ are convex and $\theta^* \in int(\Theta)$, then $\hat{\theta}_n^{\text{DSM}}$ is weakly consistent for θ^*.*

Proof. By assumption $\theta \mapsto F_\theta(x)$ is continuous. Suppose Θ is compact, taking $C = \Theta$, note

$$\begin{aligned} |F_\theta| &= \left| \left\| m^\top \nabla_x \log p_\theta \right\|_2^2 + 2\nabla \cdot \left(mm^\top \nabla \log p_\theta \right) \right| \\ &= \left| \left\| m^\top \nabla_x \log p_\theta \right\|_2^2 + 2 \left(\nabla \cdot (mm^\top) \cdot \nabla \log p_\theta + \operatorname{tr}\left[mm^\top \nabla^2 \log p_\theta \right] \right) \right| \\ &\lesssim \|m^\top\|^2 f_1^2 + 2\|\nabla \cdot (mm^\top)\| f_1 + 2\|mm^\top\|_\infty f_2 \end{aligned}$$

which is integrable, so the conditions of Lemma 2.4 [277] are satisfied so $\theta \mapsto \mathbb{Q}F_\theta$ is continuous, and $\sup_\Theta |\frac{1}{n}\sum_i^n F_\theta(X_i) - \mathbb{Q}F_\theta| \xrightarrow{p} 0$, and thus from theorem 2.1 [277] $\hat{\theta}_n^{\text{DSM}} \xrightarrow{p} \theta_*^{\text{DSM}}$. If Θ is convex, note that the sum of convex functions is convex, so $\theta \mapsto \frac{1}{n}\sum_i^n F_\theta(X_i)$ is convex, and we can follow a derivation analogous to the one in 41. ■

6.13.2.2 Asymptotic Normality of DSM

Theorem 59 (**Asymptotic Normality of DSM**). *Suppose $\mathcal{X}, \Theta$ be open subsets of $\mathbb{R}^d$ and $\mathbb{R}^m$ respectively. If **(i)** $\hat{\theta}_n^{\text{DSM}} \xrightarrow{p} \theta^*$, **(ii)** $\theta \mapsto \log p_\theta(x)$ is twice continuously differentiable on a closed ball $\bar{B}(\epsilon, \theta^*) \subset \Theta$, and*

(iii) *$\|mm^\top\| + \|\nabla_x \cdot (mm^\top)\| \leq f_1(x)$, and $\|\nabla_x \log p\| + \|\nabla_{\theta^*} \nabla_x \log p\| + \|\nabla_{\theta^*} \nabla_x \nabla_x \log p\| \leq f_2(x)$, with $f_1 f_2, f_1 f_2^2 \in L^2(\mathbb{Q})$*

(iv) *for $\theta \in \bar{B}(\epsilon, \theta^*)$ $\|\nabla_\theta \nabla_x \log p\|^2 + \|\nabla_x \log p\| \|\nabla_\theta \nabla_\theta \nabla_x \log p\| + \|\nabla_\theta \nabla_\theta \nabla_x \log p\| + \|\nabla_\theta \nabla_\theta \nabla_x \nabla_x \log p\| \leq g_1(x)$, and $f_1 g_1 \in L^1(\mathbb{Q})$,*

*and **(v)** and the information tensor is invertible at θ^*. Then*

$$\sqrt{n}\left(\hat{\theta}_n^{\text{DSM}} - \theta^* \right) \xrightarrow{d} \mathcal{N}\left(0, g^{-1}(\theta^*) \mathbb{Q}\left[\nabla_{\theta^*} F_\theta \otimes \nabla_{\theta^*} F_\theta \right] g^{-1}(\theta^*) \right)$$

Proof. From **(ii)** $\theta \mapsto F_\theta$ is twice continuously differentiable on a ball $B(\epsilon, \theta^*) \subset \Theta$. Note $\nabla_\theta \frac{1}{N}\sum_i^N F_\theta(X_i) = \frac{1}{N}\sum_i^N \nabla_\theta F_\theta(X_i)$, then $\mathbb{Q}[\nabla_\theta F_{\theta_*^{\text{DSM}}}(X_i)] = \nabla_\theta \mathbb{Q}[F_{\theta_*^{\text{DSM}}}(X_i)] = 0$. Note

$$\|\nabla_\theta F_{\theta_*^{\text{DSM}}}(x)\| \lesssim \|mm^\top\| \|\nabla_x \log p\| \|\nabla_\theta \nabla_x \log p\| + \|\nabla_x \cdot (mm^\top)\| \|\nabla_\theta \nabla_x \log p\|$$

$$
\begin{aligned}
&+ \|mm^\top\| \|\nabla_\theta \nabla_x \nabla_x \log p\| \\
&\lesssim f_1(x) f_2(x)[f_2(x) + 2].
\end{aligned}
$$

Hence $\nabla_\theta F_{\theta_*^{\text{DSM}}} \in L^2(\mathbb{Q})$, so by the CLT

$$
\sqrt{n} \nabla_\theta \frac{1}{n} \sum_i^n F_{\theta_*^{\text{DSM}}}(X_i) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{Q}\left[\nabla_\theta F_{\theta_*^{\text{DSM}}} \otimes \nabla_\theta F_{\theta_*^{\text{DSM}}}\right]\right).
$$

Now $\theta \mapsto \nabla_\theta \nabla_\theta F_\theta(x)$ is continuous on $\overline{B}(\epsilon, \theta^*)$ so we have:

$$
\begin{aligned}
\|\nabla_\theta \nabla_\theta F_\theta(x)\| \lesssim \|mm^\top\| &\left(\|\nabla_\theta \nabla_x \log p\|^2 + \|\nabla_x \log p\| \|\nabla_\theta \nabla_\theta \nabla_x \log p\|\right) \\
&+ \|\nabla \cdot (mm^\top)\| \|\nabla_\theta \nabla_\theta \nabla_x \log p\| + \|mm^\top\| \|\nabla_\theta \nabla_\theta \nabla_x \nabla_x \log p\| \\
&\lesssim f_1(x) g_1(x)
\end{aligned}
$$

Combining the above, we have that the assumptions of Lemma 2.4 [277] applied to $\overline{B}(\epsilon, \theta^*)$ hold, and $\sup_{\overline{B}(\epsilon,\theta^*)} \left|\frac{1}{n}\sum_i^n \partial_{\theta^a}\partial_{\theta^b} F_\theta|_{\theta^*}(X_i) - \mathbb{Q}\partial_{\theta^a}\partial_{\theta^b} F_\theta|_{\theta^*}\right| \xrightarrow{p} 0$. As in 42 $\mathbb{Q}\partial_{\theta^a}\partial_{\theta^b} F_\theta|_{\theta^*} = g_{ab}(\theta^*)$ is the information tensor, which is continuous at θ^* by Lemma 2.4. The result follows by theorem 3.1 [277]. ■

6.13.3 Strong Consistency and Central Limit Theorems for Exponential Families

Let $\mathcal{X}$ be an open subset of $\mathbb{R}^d$, $\Theta \subset \mathbb{R}^m$. Consider the case when the density p lies in an exponential family, i.e. $p_\theta(x) \propto \exp\left(\langle \theta, T(x)\rangle_{\mathbb{R}^m} - c(\theta)\right)\exp(b(x))$, where $\theta \in \mathbb{R}^m$ and sufficient statistic $T = (T_1, \ldots, T_m) : \mathcal{X} \to \mathbb{R}^m$. Then $\nabla T \in \Gamma(\mathcal{X}, \mathbb{R}^{m\times d})$ and $\nabla_x \log p_\theta = \nabla_x b + \theta \cdot \nabla_x T$, $\nabla_\theta \nabla_x \log p_\theta = \nabla_x T^\top$.

6.13.3.1 Strong Consistency of the Minimum Diffusion Kernel Stein Discrepancy Estimator

We consider a RKHS $\mathcal{H}^d$ of functions $f : \mathcal{X} \to \mathbb{R}^d$ with matrix kernel K. Recall the Stein kernel is

$$
\begin{aligned}
k^0 = &\nabla_x \log p \cdot m(x) K m(y)^\top \nabla_y \log p + \nabla_x \cdot (m(x)K) \cdot \nabla_y \cdot m + \operatorname{tr}\left[m(y)\nabla_y \nabla_x \cdot (m(x)K)\right] \\
&+ \nabla_y \cdot (m(y)K) \cdot m(x)^\top \nabla_x \log p + \nabla_x \cdot (m(x)K) \cdot m(y)^\top \nabla_y \log p
\end{aligned}
$$

Given a (i.i.d.) sample $X_i \sim \mathbb{Q}$, we can define an estimator using the U-statistic

$$\widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_\theta)^2 = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} k^0(X_i, X_j).$$

For the case where the density p lies in an exponential family, then $k^0 = \theta^\top A \theta + v^\top \theta + c$ where $A \in \Gamma(\mathcal{X} \times \mathcal{X}, \mathbb{R}^{m \times m}), v \in \Gamma(\mathcal{X} \times \mathcal{X}, \mathbb{R}^m)$ are given by (we set $\phi \equiv m^\top \nabla T^\top \in \Gamma(\mathcal{X}, \mathbb{R}^{d \times m})$)

$$\begin{aligned}
A &= \phi(x)^\top K(x,y) \phi(y) \\
v^\top &= \nabla_y b \cdot m(y) K(y,x) \phi(x) + \nabla_x b \cdot m(x) K(x,y) \phi(y) \\
&\quad + \nabla_x \cdot (m(x)K) \cdot \phi(y) + \nabla_y \cdot (m(y)K) \cdot \phi(x) \\
c &= \nabla_x b \cdot m(x) K(x,y) m(y)^\top \nabla_y b + \nabla_x \cdot (m(x)K) \cdot \nabla_y \cdot m + \operatorname{tr}\left[m(y) \nabla_y \nabla_x \cdot (m(x)K)\right] \\
&\quad + \nabla_y \cdot (m(y)K) \cdot m(x)^\top \nabla_x b + \nabla_x \cdot (m(x)K) \cdot m(y)^\top \nabla_y b
\end{aligned}$$

Lemma 60. *Suppose K is IPD, that ∇T has linearly independent rows, that m is invertible, and $\|\phi\|_{L^1(\mathbb{Q})} < \infty$. Then the matrix $\int_{\mathcal{X}} A\mathbb{Q} \otimes \mathbb{Q}$ is symmetric positive definite.*

Proof. The matrix $B = \int_{\mathcal{X}} A\mathbb{Q} \otimes \mathbb{Q}$ is symmetric

$$\begin{aligned}
(\int_{\mathcal{X}} A\mathbb{Q} \otimes \mathbb{Q})^\top &= \int_{\mathcal{X}} A(x,y)^\top \mathbb{Q}(\mathrm{d}x) \otimes \mathbb{Q}(\mathrm{d}y) = \int_{\mathcal{X}} \nabla_y T m(y) K(x,y)^\top m(x)^\top \nabla_x T^\top \mathbb{Q}(\mathrm{d}x) \otimes \mathbb{Q}(\mathrm{d}y) \\
&= \int_{\mathcal{X}} \nabla_y T m(y) K(y,x) m(x)^\top \nabla_x T^\top \mathbb{Q}(\mathrm{d}y) \otimes \mathbb{Q}(\mathrm{d}x) = \int_{\mathcal{X}} A\mathbb{Q} \otimes \mathbb{Q}.
\end{aligned}$$

Moreover, set $\phi \equiv m^\top \nabla T^\top$, so $A(x,y) = \phi(x)^\top K(x,y) \phi(y)$. If $v \neq 0$, then $u \equiv \phi v \neq 0$ as $\nabla T^\top$ has full column rank (i.e., the vectors $\{\nabla T_i\}$ are linearly independent) and m is invertible, and $\|\phi v\|_{L^1(\mathbb{Q})} = \int_{\mathcal{X}} \|\phi(x) v\|_1 \mathrm{d}x \le \|v\|_1 \int_{\mathcal{X}} \|\phi(x)\|_1 \mathrm{d}x < \infty$ implies $\mathrm{d}\mu_i \equiv u_i \mathrm{d}\mathbb{Q}$ is a finite signed Borel measure for each i. Clearly

$$\begin{aligned}
v^\top (\int_{\mathcal{X}} A\mathbb{Q} \otimes \mathbb{Q}) v &= \int_{\mathcal{X}} u(x)^\top K(x,y) u(y) \mathbb{Q}(\mathrm{d}x) \mathbb{Q}(\mathrm{d}y) \\
&= \int_{\mathcal{X}} K(x,y)_{ij} u_i(x) u_j(y) \mathbb{Q}(\mathrm{d}x) \mathbb{Q}(\mathrm{d}y) \\
&= \int_{\mathcal{X}} K(x,y)_{ij} \mu_i(\mathrm{d}x) \mu_j(\mathrm{d}y) \ge 0.
\end{aligned}$$

Moreover since the kernel is IPD, if this equals zero then for all i: $0 = \mu_i(C) = u_i \mathbb{Q}(C) = \phi_{ij} v_j \mathbb{Q}(C)$ for all measurable sets C, which implies $\phi v = 0$ and thus $v = 0$. ■

Theorem 61. *Suppose K is IPD with bounded derivative up to order 2, that ∇T has linearly independent rows, and m is invertible. Suppose $\|\phi\|, \|\nabla_x b\| \|m\|, \|\nabla_x m\| + \|m\| \in L^1(\mathbb{Q})$. The*

minimiser $\hat{\theta}_n^{\text{DKSD}}$ *of* $\widehat{\text{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_\theta)$ *exists eventually, and converges almost surely to the minimiser* θ^* *of* $\text{DKSD}_{K,m}(\mathbb{Q}, \mathbb{P}_\theta)$.

Proof. Let $X_i : \Omega \to \mathcal{X} \subset \mathbb{R}^d$ be independent $\mathbb{Q}$-distributed random vectors. The U-statistic $A_n \equiv \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} A(X_i, X_j)$ is symmetric semi-definite. Since $\int_{\mathcal{X}} \|A\| \mathrm{d}\mathbb{Q} \otimes \mathbb{Q} < \infty$, by theorem 1 [156] the components of A_n converge to the components of B almost surely, and since the matrix inverse is a continuous map, by the continuous mapping theorem the components of A_n^{-1} (the inverse exists eventually) converge almost surely to B^{-1}. Hence the minimiser of $\widehat{\text{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_\theta)^2 = \theta^\top A_n \theta + v_n^\top \theta + c$ where $v_n \equiv \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} v(X_i, X_j)$ exists eventually.

$$
\begin{aligned}
|A(x,y)| &\lesssim K_\infty \|\phi(x)\| \|\phi(y)\| \\
\|v\| &\lesssim K_\infty \|\nabla_y b\| \|m(y)\| \|\phi(x)\| + K_\infty \|\nabla_x b\| \|m(x)\| \|\phi(y)\| \\
&\quad + (\|\nabla_x m\| + \|m(x)\|) K_\infty \|\phi(y)\| + (\|\nabla_y m\| + \|m(y)\|) K_\infty \|\phi(x)\| \\
|c| &\lesssim K_\infty \|\nabla_x b\| \|m(x)\| \|m(y)\| \|\nabla_y b\| + K_\infty (\|\nabla_x m\| + \|m(x)\|) \|\nabla_y m\| + \\
&\quad + K_\infty \|m(y)\| (1 + \|m(x)\| + \|\nabla_x m\|) \\
&\quad + K_\infty (\|\nabla_y m\| + \|m(y)\|) \|\nabla_x m\| \|\nabla_x b\| + K_\infty (\|\nabla_x m\| + \|m(x)\|) \|\nabla_y m\| \|\nabla_y b\|
\end{aligned}
$$

and it follows from the integrability assumptions that $\mathbb{Q} \otimes \mathbb{Q} |k_\theta^0| < \infty$. Since the product and sum of random variables that converge a.s. converge a.s., we have that $\hat{\theta}_n^{\text{DKSD}} \to \theta^*$ a.s.,

$$
\hat{\theta}_n^{\text{DKSD}} = -\frac{1}{2} A_n^{-1} v_n \xrightarrow{a.s.} -\frac{1}{2} B^{-1} v = \theta^*.
$$

■

6.13.3.2 Asymptotic Normality of the DKSD Estimator

We now consider the distribution of $\sqrt{n}(\hat{\theta}_n^{\text{DKSD}} - \theta^*)$. Recall that $A \in \Gamma(\mathcal{X}, \mathbb{R}^{m \times m}), v \in \Gamma(\mathcal{X}, \mathbb{R}^m)$, and for n large enough A_n^{-1} exists a.s., and $\hat{\theta}_n^{\text{DKSD}} = -\frac{1}{2} A_n^{-1} v_n$.

Theorem 62. *Suppose* $\|\phi\|, \|\nabla_x b\| \|m\|, \|\nabla_x m\| + \|m\| \in L^2(\mathbb{Q})$. *Then the* DKSD *estimator is asymptotically normal.*

Proof. From the integrability assumptions, it follows that $v, A \in L^2(\mathbb{Q} \otimes \mathbb{Q})$, and since $\mathcal{X}$ has finite $\mathbb{Q} \otimes \mathbb{Q}$-measure, $v, A \in L^1(\mathbb{Q} \otimes \mathbb{Q})$. Assume first that $m = 1$. Hence the tuple $U_n \equiv (v_n, A_n) : \Omega \to \mathbb{R}^2$, with $\mathbb{E}[U_n] = (\int_{\mathcal{X}} v \mathbb{Q} \otimes \mathbb{Q}, \int_{\mathcal{X}} A \mathbb{Q} \otimes \mathbb{Q}) \equiv (U_1, U_2)$, is asymptotically

normal

$$\sqrt{n}(U_n - \mathbb{E}[U_n]) \xrightarrow{d} \mathcal{N}(0, 4\Sigma)$$

where, setting $v^0 = v - U_1$ and $A^0 = A - U_2$

$$\begin{aligned}\Sigma &= \mathbb{E}\left[\left(\int_{\mathcal{X}} v^0(X,y)\mathrm{d}\mathbb{Q}(y), \int_{\mathcal{X}} A^0(X,y)\mathrm{d}\mathbb{Q}(y)\right) \otimes \left(\int_{\mathcal{X}} v^0(X,y)\mathrm{d}\mathbb{Q}(y), \int_{\mathcal{X}} A^0(X,y)\mathrm{d}\mathbb{Q}(y)\right)\right] \\ &= \begin{pmatrix} \int_{\mathcal{X}} v^0(x,y)\mathrm{d}\mathbb{Q}(y) \int_{\mathcal{X}} v^0(x,z)\mathrm{d}\mathbb{Q}(z)\mathrm{d}\mathbb{Q}(x) & \int_{\mathcal{X}} v^0(x,y)\mathrm{d}\mathbb{Q}(y) \int_{\mathcal{X}} A^0(x,z)\mathrm{d}\mathbb{Q}(z)\mathrm{d}\mathbb{Q}(x) \\ \int_{\mathcal{X}} v^0(x,y)\mathrm{d}\mathbb{Q}(y) \int_{\mathcal{X}} v^0(x,z)\mathrm{d}\mathbb{Q}(z)\mathrm{d}\mathbb{Q}(x) & \int_{\mathcal{X}} A^0(x,y)\mathrm{d}\mathbb{Q}(y) \int_{\mathcal{X}} A^0(x,z)\mathrm{d}\mathbb{Q}(z)\mathrm{d}\mathbb{Q}(x) \end{pmatrix}\end{aligned}$$

Since $\hat{\theta}_n^{\mathrm{DKSD}} = g(U_n)$, $\theta^* = g(U)$ where $g(x,y) \equiv -\frac{1}{2}x/y$, we can apply the delta method which states

$$\sqrt{n}(\hat{\theta}_n^{\mathrm{DKSD}} - \theta^*) = \sqrt{n}(g(U_n) - g(U)) \xrightarrow{d} \mathcal{N}\left(0, 4\nabla g(U)\Sigma\nabla g(U)^\top\right)$$

and $\nabla g(U) = (-1/2U_2, U_1/2U_2^2)$. Now let m be arbitrary. Since $A \in L^2(\mathbb{Q})$ then setting $A^0 \equiv A - \int_{\mathcal{X}} A\mathbb{Q} \otimes \mathbb{Q}$ we find

$$\sqrt{n}(A_n - \mathbb{E}[A_n]) \xrightarrow{d} \mathcal{N}(0, 4\Sigma_1), \qquad \Sigma_1 \equiv \int_{\mathcal{X}} \left[\int_{\mathcal{X}} A^0(x,y)\mathrm{d}\mathbb{Q}(y) \otimes \int_{\mathcal{X}} A^0(x,y)\mathrm{d}\mathbb{Q}(y)\right] \mathrm{d}\mathbb{Q}(x)$$

and similarly , with $v^0 \equiv v - \int v\mathrm{d}\mathbb{Q} \otimes \mathrm{d}\mathbb{Q}$

$$\sqrt{n}(v_n - \mathbb{E}[v_n]) \xrightarrow{d} \mathcal{N}(0, 4\Sigma_2), \qquad \Sigma_2 \equiv \int_{\mathcal{X}} \left[\int_{\mathcal{X}} v^0(x,y)\mathrm{d}\mathbb{Q}(y) \otimes \int_{\mathcal{X}} v^0(x,y)\mathrm{d}\mathbb{Q}(y)\right] \mathrm{d}\mathbb{Q}(x).$$

and

$$\sqrt{n}((v_n, A_n) - \mathbb{E}[(v_n, A_n)]) \xrightarrow{d} \mathcal{N}(0, 4\Sigma)$$

where

$$\Sigma = \int_{\mathcal{X}} \left[\left(\int_{\mathcal{X}} v^0(x,y)\mathrm{d}\mathbb{Q}(y), \int_{\mathcal{X}} A^0(x,y)\mathrm{d}\mathbb{Q}(y)\right) \otimes \left(\int_{\mathcal{X}} v^0(x,y)\mathrm{d}\mathbb{Q}(y), \int_{\mathcal{X}} A^0(x,y)\mathrm{d}\mathbb{Q}(y)\right)\right] \mathrm{d}\mathbb{Q}(x).$$

Let $\mathcal{D} \equiv \mathbb{R}^m \times \mathbb{R}^{m\times m}$, which we equip with coordinates $z_{ijk} = (x_i, y_{jk})$. Consider the function $g : \mathcal{D} \to \mathbb{R}^m$, $(x,y) \mapsto -\frac{1}{2}y^{-1}x$, so $g(v_n, A_n) = \theta_n^{\mathrm{DKSD}}$. Note $\Sigma \in \mathcal{D} \times \mathcal{D}$ and $\nabla g : \mathcal{D} \to \mathrm{End}(\mathcal{D}, \mathbb{R}^m) \cong \mathbb{R}^m \times \mathcal{D}$, so that $\nabla g(U)\Sigma\nabla g(U)^\top \in \mathbb{R}^{m\times m}$. First consider the matrix inversion $h(y) = y^{-1}$, so $\nabla h(y) \in \mathbb{R}^{(m\times m)\times(m\times m)}$, and $\nabla h(y)_{(ij)(kr)} = \partial_{y^{kr}} h_{ij}$. Since $h(y)_{ij}y_{jl} = \delta_{il}$ we

have $0 = \partial_{kr}(h(y)_{ij}y_{jl}) = \partial_{kr}(h(y)_{ij})y_{jl} + h(y)_{ij}\delta_{jk}\delta_{rl} = \partial_{kr}(h(y)_{ij})y_{jl} + h(y)_{ik}\delta_{rl}$ and

$$\nabla h(y)_{(is)(kr)} = \partial_{kr}(h(y)_{ij})y_{jl}h(y)_{ls} = -h_{ik}\delta_{rl}h(y)_{ls} = -h(y)_{ik}h(y)_{rs}$$

and clearly $f : x \mapsto x$, then $\nabla f(x) = 1_{m\times m}$. Moreover

$$\partial_{y^{ab}}g_i(z) = \partial_{y^{ab}}\left(h(y)_{ij}f(x)_j\right) = \partial_{y^{ab}}(h(y)_{ij})x_j = -h(y)_{ia}h(y)_{bj}x_j, \quad \partial_{x^l}g_i(z) = h(y)_{il}$$

Then

$$\begin{aligned}(\nabla g(z)\Sigma)_{ir} = \partial_v g_i \Sigma_{vr} &= g_{i,x^l}\Sigma_{x^l r} + g_{i,y^{ab}}\Sigma_{y^{ab}r} = h(y)_{il}\Sigma_{x^l r} + \partial_{y^{ab}}(h(y)_{is})x_s\Sigma_{y^{ab}r} \\ &= h(y)_{il}\Sigma_{x^l r} - h(y)_{ia}h(y)_{bs}x_s\Sigma_{y^{ab}r},\end{aligned}$$

so

$$\begin{aligned}(\nabla g(z)\Sigma\nabla g(z)^\top)_{ic} = (\nabla g(z)\Sigma)_{ir}(\nabla g(z))_{cr} &= (\nabla g(z)\Sigma)_{ir}\partial_r g_c \\ &= h(y)_{il}\Sigma_{x^l r}\partial_r g_c - h(y)_{ia}h(y)_{bs}x_s\Sigma_{y^{ab}r}\partial_r g_c\end{aligned}$$

with

$$\begin{aligned}h(y)_{il}\Sigma_{x^l r}\partial_r g_c &= h(y)_{il}\Sigma_{x^l x^b}\partial_{x^b}g_c + h(y)_{il}\Sigma_{x^l y^{as}}\partial_{y^{as}}g_c \\ &= h(y)_{il}\Sigma_{x^l x^b}h(y)_{cb} - h(y)_{il}\Sigma_{x^l y^{as}}h_{ca}(y)h(y)_{sj}x_j\end{aligned}$$

and

$$\begin{aligned}-h(y)_{ia}h(y)_{bs}x_s\Sigma_{y^{ab}r}\partial_r g_c &= -h(y)_{ia}h(y)_{bs}x_s\left(\Sigma_{y^{ab}x^k}\partial_{x^k}g_c + \Sigma_{y^{ab}y^{ld}}\partial_{y^{ld}}g_c\right) \\ &= -h(y)_{ia}h(y)_{bs}x_s\left(\Sigma_{y^{ab}x^k}h(y)_{ck} - \Sigma_{y^{ab}y^{ld}}h(y)_{cl}h(y)_{dj}x_j\right).\end{aligned}$$

Note we have

$$\begin{aligned}\Sigma_{xx} &= \int_{\mathcal{X}}\int_{\mathcal{X}} v^0(x,y)\mathrm{d}\mathbb{Q}(y)\otimes\int_{\mathcal{X}} v^0(x,z)\mathrm{d}\mathbb{Q}(z)\mathrm{d}\mathbb{Q}(x) \equiv \int_{\mathcal{X}} T(x)\otimes T(x)\mathrm{d}\mathbb{Q}(x) \\ \Sigma_{xy} &= \int_{\mathcal{X}}\int_{\mathcal{X}} v^0(x,y)\mathrm{d}\mathbb{Q}(y)\otimes\int_{\mathcal{X}} A^0(x,z)\mathrm{d}\mathbb{Q}(z)\mathrm{d}\mathbb{Q}(x) \equiv \int_{\mathcal{X}} T(x)\otimes L(x)\mathrm{d}\mathbb{Q}(x) \\ \Sigma_{yy} &= \int_{\mathcal{X}}\int_{\mathcal{X}} A^0(x,y)\mathrm{d}\mathbb{Q}(y)\otimes\int_{\mathcal{X}} A^0(x,z)\mathrm{d}\mathbb{Q}(z)\mathrm{d}\mathbb{Q}(x) \equiv \int_{\mathcal{X}} L(x)\otimes L(x)\mathrm{d}\mathbb{Q}(x)\end{aligned}$$

then

$$4\nabla g(U_1,U_2)\Sigma\nabla g(U_1,U_2)^\top = \int_{\mathcal{X}}(U_2^{-1}T)\otimes(TU_2^{-1})\mathrm{d}\mathbb{Q}$$

$$
\begin{aligned}
&- 2\int_{\mathcal{X}} \left(U_2^{-1} L U_2^{-1} U_1\right) \otimes \left(T U_2^{-1}\right) \mathrm{d}\mathbb{Q} \\
&+ \int_{\mathcal{X}} \left(U_2^{-1} L U_2^{-1} U_1\right) \otimes \left(U_2^{-1} L U_2^{-1} U_1\right) \mathrm{d}\mathbb{Q}
\end{aligned}
$$

■

6.13.3.3 Diffusion Score Matching Asymptotics

Consider the loss function

$$
L(x,\theta) = \left\langle \nabla \log p_\theta, mm^\top \nabla \log p_\theta \right\rangle + 2\left(\nabla \cdot (mm^\top) \cdot \nabla \log p_\theta + \operatorname{tr}\left[mm^\top \nabla^2 \log p_\theta\right]\right).
$$

For the exponential family $L(x,\theta) = \theta^\top A\theta + v^\top\theta + c$, where (we set $S = mm^\top$)

$$
\begin{aligned}
A &= \nabla T S \nabla T^\top \\
v^\top &= 2\nabla b \cdot S \nabla T^\top + 2\nabla \cdot S \cdot \nabla T^\top + 2\operatorname{tr}\left[S\nabla^2 T_i\right] e_i \\
c &= \nabla b \cdot S \nabla b + 2\nabla \cdot S \cdot \nabla b + 2\operatorname{tr}[S\nabla\nabla b].
\end{aligned}
$$

Theorem 63. *Suppose m is invertible and $\{\nabla T_i\}$ are linearly independent. Then if $A, v \in L^1(\mathbb{Q})$, $\hat{\theta}_n^{\mathrm{DSM}}$ eventually exists and is strongly consistent. If we also have $A, v \in L^2(\mathbb{Q})$, then $\hat{\theta}_n^{\mathrm{DSM}}$ is asymptotically normal.*

Proof. Let $\mathcal{M} \equiv \int A\mathrm{d}\mathbb{Q}$, $H \equiv \int v\mathrm{d}\mathbb{Q}$. If $A = \nabla T mm^\top \nabla T^\top = \nabla T m (\nabla T m)^\top$ so $\operatorname{rank} A = \operatorname{rank}(\nabla T m(\nabla T m)^\top) = \operatorname{rank}(\nabla T m) = \operatorname{rank}(\nabla T) = \operatorname{rank}(\nabla T^\top)$ if m is invertible. So if the vectors $\{\nabla T_i\}$ are linearly independent, then $\nabla T^\top$ has full column rank. Then A it is symmetric positive (strictly) definite and the minimum of $L(\theta) \equiv \int L(x,\theta)\mathrm{d}\mathbb{Q}(x)$ is $\theta^* = -\frac{1}{2}\mathcal{M}^{-1}H$ which for sufficiently large n can be estimated by the random variable $\hat{\theta}_n^{\mathrm{DSM}} \equiv -\frac{1}{2}\mathcal{M}_n^{-1}H_n$ which converges a.s. to θ.

We consider the tuple $U_n \equiv (H_n, \mathcal{M}_n)$, so $\mathbb{E}[U_n] = (H, \mathcal{M})$. Since $A, v \in L^2(\mathbb{Q})$, then

$$
\sqrt{n}\left(U_n - (H,\mathcal{M})\right) \xrightarrow{d} \mathcal{N}(0,\Gamma)
$$

where, setting $v^0 = v - H$, $A^0 = A - \mathcal{M}$

$$
\Gamma = \mathbb{E}\left[(v^0, A^0) \otimes (v^0, A^0)\right].
$$

Let $\mathcal{D} \equiv \mathbb{R}^m \times \mathbb{R}^{m\times m}$, and consider $g : \mathcal{D} \to \mathbb{R}^m$, defined by $g(x,y) = -\frac{1}{2}y^{-1}x$. Using the Delta

method

$$\sqrt{n}(\hat{\theta}_n^{\mathrm{DSM}} - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, 4\nabla g(H,\mathcal{M})\Gamma\nabla g(H,\mathcal{M})^\top\right)$$

where, proceeding as in 6.13.3.2, we find

$$\begin{aligned} 4\nabla g(H,\mathcal{M})\Gamma\nabla g(H,\mathcal{M})^\top = &\int_{\mathcal{X}} (\mathcal{M}^{-1}v^0) \otimes (v^0\mathcal{M}^{-1})\mathrm{d}\mathbb{Q} \\ &- 2\int_{\mathcal{X}} \left(\mathcal{M}^{-1}A^0\mathcal{M}^{-1}H\right) \otimes \left(v^0\mathcal{M}^{-1}\right)\mathrm{d}\mathbb{Q} \\ &+ \int_{\mathcal{X}} \left(\mathcal{M}^{-1}A^0\mathcal{M}^{-1}H\right) \otimes \left(\mathcal{M}^{-1}A^0\mathcal{M}^{-1}H\right)\mathrm{d}\mathbb{Q} \end{aligned}$$

■

6.14 Proofs of Robustness of Minimum Stein Discrepancy Estimators

In this section, we provide conditions on the Stein operator (and Stein class) to obtain robust estimators in the context of DKSD and DSM. In particular we prove 45 and derive the influence function of DSM.

6.14.1 Robustness of Diffusion Kernel Stein Discrepancy

Let $T : \mathcal{P}_\Theta \to \Theta$ with $T(\mathbb{P}) = \mathrm{argmin}_\Theta \, \mathrm{DKSD}_{K,m}(\mathbb{P}\|\mathbb{P}_\theta)$ be defined by $\mathrm{IF}(z,\mathbb{Q}) \equiv \lim_{t\to 0}(T(\mathbb{Q}+t(\delta_z-\mathbb{Q})) - T(\mathbb{Q}))/t$. Denote $\mathbb{Q}_t = \mathbb{Q} + t(\delta_z - \mathbb{Q})$, $\theta_t = T(\mathbb{Q}_t)$, $\theta_0 = T(\mathbb{Q})$. Note that by the first order optimality condition:

$$\nabla_\theta \int_{\mathcal{X}}\int_{\mathcal{X}} k^0 \mathbb{Q}_t \otimes \mathbb{Q}_t|_{\theta_t} = \nabla_{\theta_t} \, \mathrm{DKSD}_{K,m}(\mathbb{Q}_t\|\mathbb{P}_\theta) = 0.$$

By the MVT, there exists $\bar{\theta}$ on the line joining θ_0 and θ_t for which

$$0 = \int_{\mathcal{X}}\int_{\mathcal{X}} \nabla_\theta k^0|_{\theta_0}\mathbb{Q}_t \otimes \mathbb{Q}_t + \int_{\mathcal{X}}\int_{\mathcal{X}} \nabla_\theta\nabla_\theta k^0|_{\bar{\theta}}\mathbb{Q}_t \otimes \mathbb{Q}_t(\theta_t - \theta_0).$$

Expanding

$$\mathbb{Q}_t \otimes \mathbb{Q}_t \nabla_\theta k^0|_{\theta_0} = t^2(\delta_z - \mathbb{Q}) \otimes (\delta_z - \mathbb{Q}) \nabla_\theta k^0|_{\theta_0} + 2t\mathbb{Q}_y \nabla_\theta k^0|_{\theta_0}(z, y)$$

where we have used the optimality condition. On the other hand

$$\mathbb{Q}_t \otimes \mathbb{Q}_t \nabla_\theta \nabla_\theta k^0|_{\bar{\theta}} = (1-2t)\mathbb{Q} \otimes \mathbb{Q} \nabla_\theta \nabla_\theta k^0|_{\bar{\theta}} + t^2(\delta_z - \mathbb{Q}) \otimes (\delta_z - \mathbb{Q}) \nabla_\theta \nabla_\theta k^0|_{\bar{\theta}} + 2t\mathbb{Q}_y \nabla_\theta \nabla_\theta k^0|_{\bar{\theta}}(z, y).$$

Hence

$$\mathbb{Q}_y \nabla_\theta k^0|_{\theta_0}(z, y) = \tfrac{1}{2} \left((1-2t)\mathbb{Q} \otimes \mathbb{Q} \nabla_\theta \nabla_\theta k^0|_{\bar{\theta}} + 2t\mathbb{Q}_y \nabla_\theta \nabla_\theta k^0|_{\bar{\theta}}(z, y) \right) \frac{\theta_t - \theta_0}{t} + O(t),$$

and taking the limit $t \to 0$, $\bar{\theta} \to \theta_0$ and using a derivation as in the proof of 42

$$\mathbb{Q}_y \nabla_\theta k^0|_{\theta_0}(z, y) = \tfrac{1}{2} \int_{\mathcal{X}} \int_{\mathcal{X}} \nabla_\theta \nabla_\theta k^0|_{\theta_0} \mathrm{d}\mathbb{Q} \otimes \mathrm{d}\mathbb{Q}\, \mathrm{IF}(z, \mathbb{Q}) = g(\theta_0)\, \mathrm{IF}(z, \mathbb{Q})$$

hence the influence function is given by

$$\mathrm{IF}(z, \mathbb{Q}) = g(\theta_0)^{-1} \int_{\mathcal{X}} \nabla_\theta k^0|_{\theta_0}(z, y) \mathrm{d}\mathbb{Q}(y).$$

We aim to show the estimator is B-robust, that is $z \mapsto \|\mathrm{IF}(z, \mathbb{Q})\|$ is bounded. Suppose that the additional assumptions hold. Then there exists a function c such that $\int \langle s_p(x), K(x, y) \nabla_{\theta_0} s_p(y) \rangle \mathbb{Q}(dy) \leq \|s_p(x)\| c(x)$ which is bounded in $x \in \mathcal{X}$. Following a similar argument, and using the assumptions, a similar limit will hold for all terms in $\int \nabla_{\theta_0} k^0(z, y) \mathrm{d}\mathbb{Q}(y)$. It follows that $\sup_{z \in \mathcal{X}} \|\mathrm{IF}(z, \mathbb{Q})\| < \infty$.

6.14.2 Robustness of Diffusion Score Matching

The scoring rule $S : \mathcal{X} \times \mathcal{P}_{\mathcal{X}} \to \mathbb{R}$ of DSM is

$$S(x, \mathbb{P}_\theta) \equiv \tfrac{1}{2} \left\| m^\top \nabla_x \log p_\theta \right\|_2^2 + \nabla \cdot \left(m m^\top \nabla \log p_\theta \right)(x)$$

Indeed the proof of 38 we have

$$\int_{\mathcal{X}} \left\| m^\top \nabla \log q \right\|^2 \mathrm{d}\mathbb{Q} = - \int_{\mathcal{X}} \nabla \cdot \left(m m^\top \nabla \log q \right) \mathrm{d}\mathbb{Q}.$$

which implies $\mathbb{Q}S(\cdot, \mathbb{Q}) = -\frac{1}{2}\int_{\mathcal{X}} \left\|m^\top \nabla \log q\right\|^2 \mathrm{d}\mathbb{Q}$, so

$$\begin{aligned}\mathbb{Q}S(\cdot, \mathbb{P}_\theta) - \mathbb{Q}S(\cdot, \mathbb{Q}) &= \int_{\mathcal{X}} \left(\tfrac{1}{2}\left\|m^\top \nabla_x \log p_\theta\right\|_2^2 + \tfrac{1}{2}\left\|m^\top \nabla \log q\right\|^2 + \nabla \cdot \left(mm^\top \nabla \log p_\theta\right)\right) \mathrm{d}\mathbb{Q} \\ &= \mathrm{DSM}_m(\mathbb{Q} \| \mathbb{P}_\theta).\end{aligned}$$

From 4.2 [95] the influence function is then $\mathrm{IF}(x, \mathbb{P}_\theta) = g_{\mathrm{DSM}}(\theta)^{-1} s(x, \theta)$, where

$$\begin{aligned}s(x,\theta) \equiv \nabla_\theta S(x,\theta) &= \tfrac{1}{2}\nabla_\theta \|m^\top \nabla_x \log p_\theta\|_2^2 + \nabla_\theta \nabla_x \cdot \left(mm^\top \nabla_x \log p_\theta\right) \\ &= \tfrac{1}{2}\nabla_\theta \|m^\top \nabla_x \log p_\theta\|_2^2 + \nabla_\theta \left(\left\langle \nabla_x \cdot (mm^\top), \nabla \log p_\theta \right\rangle + \mathrm{tr}\left[mm^\top \nabla_x^2 \log p_\theta\right]\right) \\ &= \nabla_x \nabla_\theta \log p_\theta mm^\top \nabla_x \log p_\theta + (\nabla_x \nabla_\theta \log p_\theta) \nabla_x \cdot (mm^\top) + \mathrm{Tr}[mm^\top \nabla_x \nabla_x] \nabla_\theta \log p_\theta\end{aligned}$$

and where $g_{\mathrm{DSM}}(\theta) \equiv \mathbb{P}_\theta \nabla_\theta \nabla_\theta S(\cdot, \theta)$ is the information metric associated with DSM. Hence the estimator is bias-robust iff $x \mapsto s(x, \theta_*^{\mathrm{DSM}})$ is bounded.

Chapter 7

Conclusion

In this thesis we have constructed a deep geometric understanding of many statistical methods, in particular the theory of smooth measures and score-based samplers, and we have introduced many ideas from geometric mechanics which will hopefully motivate fruitful collaborations between geometers and statisticians.

We have developed a fully geometric theory of HMC, explained how to implement HMC with embedding and submersions in full generality, and showed that the theory of mechanics with symmetries provided a unified framework to implement many HMC samplers in statistics and lattice QCD, where distributions with symmetries (or on shape spaces) are common. The comprehensive HMC theory clarifies several recurring misconceptions, such as the role of Lagrangian mechanics, and the implementation of HMC on manifolds. Moreover we have carefully analysed the bracket geometry of HMC, to explain how to devise more generalised score-based dynamical systems which maintain some of the benefits of HMC. Of course the most important results are the construction of the canonical geometry of smooth measures and the canonical characterisation of measure-preserving diffusions, measure-preserving Nambu mechanics, and measure-constraints preserving flows. These results provide the starting point in the development of new diffusion-based Markov chains, and the formalism allows to easily adapt theorems from differential geometry to the statistical context, and show the importance of (co)homology in the geometry of statistics. As an immediate corollary of our formalism, we have improved and extended to arbitrary manifolds the celebrated complete recipe derived in [242], with no other assumption than the smoothness and non-degeneracy of the target measure, thus creating a link between MCMC methods and one of the most important concepts of theoretical physics: the gauge field or vector potential. In particular, the rotationnel introduced in this thesis will have an important impact on many areas of statistics and machine learning, since it adapts the exterior derivative to the distribution of interest, and shows that while physics is geometrised

via twisted differential forms and connections, in statistics multi-vector fields are bound to play a central role. In particular we have shown that, when combined with the bracket formalism, it naturally allowed to generalise Stein operators, and score-matching to arbitrary manifolds, without relying on Riemannian metrics and generators of diffusions. We have constructed minimum maximum mean discrepancy estimators for likelihood-free models with many theoretical guarantees, and shown that minimum Stein discrepancy estimators improved on some of the best methods for the estimation of measures known only through samples through unnormalised statistical models, in particular in terms of robustness.

Concerning HMC, future work should look at applications of Poisson-Hamiltonian systems as well as understanding how to use the electromagnetic symplectic structure to improve ergodicity of the chain, by adding a magnetic force on appropriate regions of the sample space. More efficient geometric integrators for HMC on holonomic systems are of interest for both statistics and molecular dynamics. It would be interesting to analyse the generalisation of RATTLE [259], as well as integrators on submanifolds obtained by reducing a system with symmetries on the unconstrained embedded spaces. Concerning diffusions, the geometrisation of measure-preserving diffusions offers limitless opportunities. Its extension to Hilbert spaces and Hilbert manifolds would be particularly useful. A theory of symmetries and reduction by symmetries for general brackets would also be interesting, since any result would have immediate consequences in both Stein operators, HMC, measure-preserving diffusions, and any statistical method based on brackets. The approach of adapting the properties of the bracket to the statistical application of interest seems very powerful. In this thesis we have mainly focused on the construction of systems with measure/density-preserving properties. Mechanics also plays a central role in optimisation, where dissipative systems have recently received considerable attention. It would be interesting to consider the application of the bracket formalism to the problem of optimisation, and analyse which bracket properties lead to efficient optimisation algorithms. A good place to start would be the metriplectic formalism, which combines a conservative and dissipative bracket in a compatible manner.

Bibliography

[1] Assyr Abdulle, Grigorios A Pavliotis, and Gilles Vilmart. Accelerated convergence to equilibrium and reduced asymptotic variance for Langevin dynamics using Stratonovich perturbations. *Comptes Rendus Mathematique*, 357(4):349–354, 2019. 4.4.3

[2] Ralph Abraham and Jerrold E. Marsden. *Foundations of Mechanics*. American Mathematical Society, second edition, May 2008. with the assistance of Tudor Ratiu and Richard Cushman. 3, 3.1.1, 3.2.2.1, 3.2.2.2, 3.4.1, 3.5.3, 3.8.1, 3.8.1, 3.8.1.1, 3.8.1.2, 3.8.2, 3.9, 3.9.1.1, A.5.6.1, A.5.6.2, A.5.6.3, A.5.6.3, A.5.6.4

[3] Ralph Abraham, Jerrold E Marsden, and Tudor Ratiu. *Manifolds, tensor analysis, and applications*, volume 75. Springer Science & Business Media, 2012. 3.2.1, 3.2.2.1, 3.3, 2, 4.2.1, 4.2.1, 4.3.3, 4.4.5, A.5.1

[4] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Applicandae Mathematica*, 80(2):199–220, 2004. 3.7.1

[5] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009. A.5.2

[6] P-A Absil and Jérôme Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012. A.5.2

[7] Elena Akhmatskaya, Nawaf Bou-Rabee, and Sebastian Reich. A comparison of generalized hybrid Monte Carlo methods with and without momentum flip. *Journal of Computational Physics*, 228(6):2256–2265, 2009. 3.4.2.3

[8] Elena Akhmatskaya and Sebastian Reich. GSHMC: An efficient method for molecular simulation. *Journal of Computational Physics*, 227(10):4934–4954, 2008. A.4.2

[9] Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 15(1):3563–3593, 2014. A.4.7

[10] Dmitri Alekseevsky and Andreas Arvanitoyeorgos. Riemannian flag manifolds with homogeneous geodesics. *Transactions of the American Mathematical Society*, 359(8):3769–3789, 2007. 3.1.3.2, 3.9.2.2

[11] Dmitri V. Alekseevsky, Janusz Grabowski, Giuseppe Marmo, and Peter W. Michor. Poisson structures on the cotangent bundle of a Lie group or a principle bundle and their reductions. *Journal of Mathematical Physics*, 35(9):4909–4927, 1994. 1.1, 3.3, 3.3.1, 3.9.1.1

[12] S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998. 5.2, 5.2

[13] Esteban Andruchow, Gabriel Larotonda, Lazaro Recht, and Alejandro Varela. The left invariant metric in the general linear group. *Journal of Geometry and Physics*, 86:241–257, 2014. 3.9.2

[14] S Aoki, R Burkhalter, M Fukugita, S Hashimoto, KI Ishikawa, N Ishizuka, Y Iwasaki, K Kanaya, T Kaneko, Y Kuramashi, et al. Polynomial hybrid Monte Carlo algorithm for lattice QCD with an odd number of flavors. *Physical review D*, 65(9):094507, 2002. A.4.1

[15] Cédric Archambeau, Manfred Opper, Yuan Shen, Dan Cornford, and John S Shawe-Taylor. Variational inference for diffusion processes. In *Advances in Neural Information Processing Systems*, pages 17–24, 2008. 2.2

[16] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223, 2017. A.4.6

[17] Alexis Arnaudon, Alessandro Barp, and So Takao. Irreversible Langevin MCMC on Lie groups. *arXiv preprint arXiv:1903.08939*, 2019. 1.1, 3.4.2.3, 4, 4.5, 4.5.2.2, 4.5.2.2

[18] Marc Arnaudon, Anton Thalmaier, et al. The differentiation of hypoelliptic diffusion semigroups. *Illinois Journal of Mathematics*, 54(4):1285–1311, 2010. 18, A.7.1

[19] Vladimir Igorevich Arnol'd. *Mathematical methods of classical mechanics*, volume 60. Springer Science & Business Media, 2013. 3.5.2

[20] Andreas Arvanitoyeorgos. Homogeneous manifolds whose geodesics are orbits. recent results and some open problems. *arXiv preprint arXiv:1706.09618*, 2017. 3.1.3.2

[21] Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. *Information geometry*, volume 64. Springer, 2017. 3.6.1, 5.1.2, 5.1.3

[22] Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. Parametrized measure models. In *Information Geometry*, pages 121–184. Springer, 2017. 5.1.1

[23] Edoardo Ballico, Alessandra Bernardi, Iacopo Carusotto, Sonia Mazzucchi, and Valter Moretti. *Quantum Physics and Geometry*. Springer, 2019. 4.2.2

[24] Andrew D Barbour. Stein's method and Poisson process convergence. *Journal of Applied Probability*, 25(A):175–184, 1988. 5.4.2, 6.9

[25] Ole E Barndorff-Nielsen and Peter E Jupp. Statistics, yokes and symplectic geometry. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 6, pages 389–427, 1997. A.5.6.5

[26] Alessandro Barp. Hamiltonian Monte Carlo on Lie groups and constrained mechanics on homogeneous manifolds. *arXiv preprint arXiv:1903.04662*, 2019. 3.3.1

[27] Alessandro Barp, Francois-Xavier Briol, Andrew Duncan, Mark Girolami, and Lester Mackey. Minimum Stein discrepancy estimators. In *Advances in Neural Information Processing Systems*, pages 12964–12976, 2019. 6, 6.8

[28] Alessandro Barp, Anthony Kennedy, and Mark Girolami. Hamiltonian Monte Carlo on symmetric and homogeneous spaces via symplectic reduction. *arXiv preprint arXiv:1903.02699*, 2019. 3.8.1.2, 3.8.3, 3.9.1.3

[29] Alessandro Barp, Chris Oates, Emilio Porcu, Mark Girolami, et al. A Riemannian-Stein kernel method. *arXiv preprint arXiv:1810.04946*, 2018. 1.1, 5, 6.9

[30] Federico Bassetti, Antonella Bodini, and Eugenio Regazzini. On minimum Kantorovich distance estimators. *Statistics & probability letters*, 76(12):1298–1302, 2006. 2.2.1

[31] GG Batrouni, H Kawai, and Pietro Rossi. Coordinate-independent formulation of the Langevin equation. *Journal of mathematical physics*, 27(6):1646–1648, 1986. 4.6.3

[32] Denis R Bell. Stochastic differential equations and hypoelliptic operators. In *Real and stochastic analysis*, pages 9–42. Springer, 2004. 18

[33] Denis R Bell, Salah-Eldin A Mohammed, et al. An extension of Hormander's theorem for infinitely degenerate second-order operators. *Duke Mathematical Journal*, 78(3):453–476, 1995. 18

[34] Giancarlo Benettin and Antonio Giorgilli. On the Hamiltonian interpolation of near-to-the identity symplectic mappings with application to symplectic integration algorithms. *Journal of Statistical Physics*, 74(5-6):1117–1143, 1994. 3.5.4.1, 3.5.4.3

[35] Berestovskiĭ. Geodesics and shortest arcs of a special sub-Riemannian metric on the Lie group so (3). 3.9.2

[36] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011. 5.3

[37] Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(2):235–269, 2019. A.8

[38] Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 2019. 2.2.1

[39] Alexandros Beskos, Natesh Pillai, Gareth Roberts, Jesus-Maria Sanz-Serna, Andrew Stuart, et al. Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, 19(5A):1501–1534, 2013. 2.1.3

[40] Alexandros Beskos, Frank J Pinski, Jesús Maria Sanz-Serna, and Andrew M Stuart. Hybrid Monte Carlo on Hilbert spaces. *Stochastic Processes and their Applications*, 121(10):2201–2230, 2011. A.4.4

[41] Michael Betancourt. A general metric for Riemannian manifold Hamiltonian Monte Carlo. In *Geometric science of information*, pages 327–334. Springer, 2013. 2.1.2

[42] Michael Betancourt. The fundamental incompatibility of scalable Hamiltonian Monte Carlo and naive data subsampling. In *International Conference on Machine Learning*, pages 533–540, 2015. A.4.3

[43] Michael Betancourt. Identifying the optimal integration time in Hamiltonian Monte Carlo. *arXiv preprint arXiv:1601.00225*, 2016. A.4.5

[44] Michael Betancourt, Simon Byrne, Sam Livingstone, Mark Girolami, et al. The geometric foundations of Hamiltonian Monte Carlo. *Bernoulli*, 23(4A):2257–2298, 2017. 3.3

[45] MJ Betancourt. Adiabatic Monte Carlo. *arXiv preprint arXiv:1405.3489*, 2014. A.5.6.1

[46] MJ Betancourt, Simon Byrne, and Mark Girolami. Optimizing the integrator step size for Hamiltonian Monte Carlo. *arXiv preprint arXiv:1411.6669*, 2014. 4

[47] Jérémie Bigot, Elsa Cazelles, and Nicolas Papadakis. Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications. 2019. A.4.8

[48] Jean-Michel Bismut. Martingales, the Malliavin calculus and hypoellipticity under general Hörmander's conditions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 56(4):469–505, 1981. 18

[49] Sergio Blanes, Fernando Casas, and JM Sanz-Serna. Numerical integrators for the Hybrid Monte Carlo method. *SIAM Journal on Scientific Computing*, 36(4):A1556–A1580, 2014. 2.1.3

[50] Vladimir I Bogachev. *Measure theory*, volume 1. Springer Science & Business Media, 2007. 6.9

[51] S. Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013. 5.2

[52] Raoul Bott and Loring W Tu. *Differential forms in algebraic topology*, volume 82. Springer Science & Business Media, 2013. 4.2.1

[53] Nawaf Bou-Rabee, Aleksandar Donev, and Eric Vanden-Eijnden. Metropolis integration schemes for self-adjoint diffusions. *Multiscale Modeling & Simulation*, 12(2):781–831, 2014. 17

[54] Nawaf Bou-Rabee and Jesús María Sanz-Serna. Geometric integrators and the Hamiltonian Monte Carlo method. *Acta Numerica*, 27:113–206, 2018. 2.1.3

[55] Nawaf Bou-Rabee, Jesús María Sanz-Serna, et al. Randomized Hamiltonian Monte Carlo. *The Annals of Applied Probability*, 27(4):2159–2194, 2017. A.4.5

[56] Nawaf Bou-Rabee and Eric Vanden-Eijnden. Pathwise accuracy and ergodicity of metropolized integrators for SDEs. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 63(5):655–696, 2010. 3.4.2.3

[57] Nawaf Bou-Rabee and Eric Vanden-Eijnden. A patch that imparts unconditional stability to explicit integrators for Langevin-like equations. *Journal of Computational Physics*, 231(6):2565–2580, 2012. 3.4.2.3

[58] Mohamed Boucetta. Poisson manifolds with compatible pseudo-metric and pseudo-Riemannian Lie algebras. *Differential Geometry and its Applications*, 20(3):279–291, 2004. 4.3.1.6

[59] Nicolas Bourbaki. *Variétés différentielles et analytiques: Fascicule de résultats*. Springer Science & Business Media, 2007. 3.2.2.1

[60] Alessandro Bravetti, Hans Cruz, and Diego Tapias. Contact Hamiltonian mechanics. *Annals of Physics*, 376:17–39, 2017. 4.3.2

[61] Alessandro Bravetti and Diego Tapias. Liouville's theorem and the canonical measure for nonconservative systems from contact geometry. *Journal of Physics A: Mathematical and Theoretical*, 48(24):245001, 2015. 3.2.1

[62] F-X. Briol, A. Barp, A. B. Duncan, and M. Girolami. Statistical inference for generative models with maximum mean discrepancy. *arXiv:1906.05944*, 2019. 1.1, 2.2.1, 5, 5.5, A.8.3

[63] Marcus Brubaker, Mathieu Salzmann, and Raquel Urtasun. A family of MCMC methods on implicitly defined manifolds. In *Artificial intelligence and statistics*, pages 161–172, 2012. 3.5.2.1

[64] Robert L Bryant. An introduction to Lie groups and symplectic geometry, 1991. 3.9.3.3

[65] Robert L Bryant and NC Durham. An introduction to Lie groups and symplectic geometry. *Geometry and quantum field theory*, 1:321–347, 1995. 4.6.4

[66] Simon Byrne and Mark Girolami. Geodesic Monte Carlo on Embedded Manifolds. *Scandinavian Journal of Statistics, Theory and Applications*, 2013. 1.1, 3, 3.1, 3.6, 3.6.1, 3.6.2, 3.6.4, 3.6.4, 3.6.5, 3.8

[67] Cédric M Campos and JM Sanz-Serna. Extra chance generalized hybrid Monte Carlo. *Journal of Computational Physics*, 281:365–374, 2015. 3.4.2.2

[68] Cédric M Campos and JM Sanz-Serna. Palindromic 3-stage splitting integrators, a roadmap. *Journal of Computational Physics*, 346:340–355, 2017. 2.1.3

[69] Massimo Campostrini and Paolo Rossi. A comparison of numerical algorithms for dynamical fermions. *Nuclear Physics B*, 329(3):753–764, 1990. 2.1.3

[70] Eric Cances, Frédéric Legoll, and Gabriel Stoltz. Theoretical and numerical comparison of some sampling methods for molecular dynamics. *ESAIM: Mathematical Modelling and Numerical Analysis*, 41(2):351–389, 2007. 2.1, 2.1

[71] JF Cariñena, F Falceto, J Grabowski, and MF Rañada. Geometry of Lie integrability by quadratures. *Journal of Physics A: Mathematical and Theoretical*, 48(21):215206, 2015. 4.6.4

[72] Gilles Carron. L^2 harmonics forms on non compact manifolds. *arXiv preprint arXiv:0704.3194*, 2007. 2

[73] EA Carter, Giovanni Ciccotti, James T Hynes, and Raymond Kapral. Constrained reaction coordinate dynamics for the simulation of rare events. *Chemical Physics Letters*, 156(5):472–477, 1989. 3.2.2.1

[74] Patrick Cattiaux and Laurent Mesnager. Hypoelliptic non-homogeneous diffusions. *Probability Theory and Related Fields*, 123(4):453–483, 2002. 18

[75] Elena Celledoni, Håkon Marthinsen, and Brynjulf Owren. An introduction to Lie group integrators–basics, new developments and applications. *Journal of Computational Physics*, 257:1040–1061, 2014. A.5.2

[76] C. Ceylan and M. U. Gutmann. Conditional noise-contrastive estimation of unnormalised models. *arXiv:1806.03664*, 2018. 6

[77] Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *International conference on machine learning*, pages 1683–1691, 2014. A.4.3

[78] W. Y. Chen, A. Barp, F-X. Briol, J. Gorham, M. Girolami, L. Mackey, and C. J. Oates. Stein point Markov Chain Monte Carlo. In *International Conference on Machine Learning, PMLR 97*, pages 1011–1021, 2019. 6.3

[79] W. Y. Chen, L. Mackey, J. Gorham, F-X. Briol, and C. J. Oates. Stein points. In *Proceedings of the International Conference on Machine Learning, PMLR 80:843-852*, 2018. 5.5, 6.3

[80] Wilson Ye Chen, Alessandro Barp, François-Xavier Briol, Jackson Gorham, Mark Girolami, Lester Mackey, Chris Oates, et al. Stein point Markov Chain Monte Carlo. *arXiv preprint arXiv:1905.03673*, 2019. 1.1, 5, 5.5

[81] Yifan Chen and Wuchen Li. Natural gradient in Wasserstein statistical manifold. *arXiv preprint arXiv:1805.08380*, 2018. 5.1.2, 5.2

[82] Youngmin Cho and Lawrence K Saul. Kernel methods for deep learning. In *Advances in neural information processing systems*, pages 342–350, 2009. 5.3

[83] K. Chwialkowski, H. Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning*, pages 2606–2615, 2016. 6.3, 6.3

[84] Daniel Cibotaru and Jorge de Lira. A note on the area and coarea formulas for general volume densities and some applications. *Manuscripta Mathematica*, 149(3-4):471–506, 2016. 3.2.2.1, 3.3

[85] MA Clark and AD Kennedy. Asymptotics of fixed point distributions for inexact Monte Carlo algorithms. *Physical Review D*, 76(7):074508, 2007. 2.1.3, 3.5.4.2

[86] Michael A Clark, Bálint Joó, Anthony D Kennedy, and Paolo J Silva. Improving dynamical lattice QCD simulations through integrator tuning using Poisson brackets and a force-gradient integrator. *Physical Review D*, 84(7):071502, 2011. A.4.1

[87] Michael A Clark, Anthony D Kennedy, and PJ Silva. Tuning HMC using Poisson brackets. *arXiv preprint arXiv:0810.1315*, 2008. 3.9

[88] Adam D Cobb, Atılım Güneş Baydin, Andrew Markham, and Stephen J Roberts. Introducing an explicit symplectic integration scheme for Riemannian manifold Hamiltonian Monte Carlo. *arXiv preprint arXiv:1910.06243*, 2019. 5.1.4

[89] Luca Consolini, Alessandro Costalunga, and Manfredi Maggiore. A coordinate-free theory of virtual holonomic constraints. *arXiv preprint arXiv:1709.07726*, 2017. 3.6.2.1

[90] Michael Creutz and Andreas Gocksch. Higher-order hybrid Monte Carlo algorithms. *Physical Review Letters*, 63(1):9, 1989. 2.1.3

[91] Ana Bela Cruzeiro, Darryl D Holm, and Tudor S Ratiu. Momentum maps and stochastic clebsch action principles. *Communications in Mathematical Physics*, 357(2):873–912, 2018. 4.2.2.3, 4.3.1.3, 4.4.1, 4.4.1

[92] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013. A.4.8

[93] Ana Cannas Da Silva. *Lectures on symplectic geometry*, volume 3575. Springer, 2001. 3.5.3.2, 4.2.2.3, 4.3.1.1

[94] A. P. Dawid, M. Musio, and L. Ventura. Minimum scoring rule inference. *Scandinavian Journal of Statistics*, 43(1):123–138, 2016. 6.1

[95] A. P. Dawid and Monica Musio. Theory and applications of proper scoring rules. *Metron*, 72(2):169–183, 2014. 6.14.2

[96] Georges De Rham. *Sur l'analysis situs des variétés à n dimensions*, volume 1305. Gauthier-Villars, 1931. 4.6.3

[97] Gerardo F Torres Del Castillo. *Differentiable Manifolds: A Theoretical Physics Approach*. Springer Science & Business Media, 2011. 3.9.1.1

[98] G. Detommaso, T. Cui, Y. Marzouk, A. Spantini, and R. Scheichl. A Stein variational newton method. In *Advances in Neural Information Processing Systems 31*, pages 9169–9179. 2018. 6.3

[99] Mingnan Ding, Zhanchun Tu, and Xiangjun Xing. Covariant formulation of nonlinear Langevin theory with multiplicative Gaussian white noises. *arXiv preprint arXiv:2007.16131*, 2020. 4.6.3

[100] P Dobson, I Fursov, G Lord, and M Ottobre. Reversible and non-reversible Markov Chain Monte Carlo algorithms for reservoir simulation problems. *arXiv preprint arXiv:1903.06960*, 2019. 3.4.2.3, 4.5.2.2

[101] J-P Dufour and Abdeljalil Haraki. Rotationnnels et structures de Poisson quadratiques. *Comptes rendus de l'Académie des sciences. Série 1, Mathématique*, 312(1):137–140, 1991. 4.2.1, 4.3.1.6

[102] Jean-Paul Dufour and Nguyen Tien Zung. *Poisson structures and their normal forms*, volume 242. Springer Science & Business Media, 2006. 3, 4.2.2.2, 6, 4.3.1.3, 28

[103] AB Duncan, GA Pavliotis, and KC Zygalakis. Nonreversible Langevin samplers: Splitting schemes, analysis and implementation. *arXiv preprint arXiv:1701.04247*, 2017. 3.4.2.3, 4.4.3

[104] Andrew B Duncan, Tony Lelievre, and GA Pavliotis. Variance reduction using nonreversible Langevin samplers. *Journal of statistical physics*, 163(3):457–491, 2016. 3.4.2.3

[105] Alain Durmus, Eric Moulines, and Eero Saksman. On the convergence of Hamiltonian Monte Carlo. *arXiv preprint arXiv:1705.00166*, 2017. A.4.5

[106] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Uncertainty in Artificial Intelligence*, 2015. A.8.3.1

[107] Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015. 5.5, A.4.6

[108] Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. 3.6.4

[109] Denis J Evans and Gary Morriss. *Statistical mechanics of nonequilibrium liquids*. Cambridge University Press, 2008. 3.2.1

[110] Lawrence Craig Evans and Ronald F Gariepy. *Measure theory and fine properties of functions*. Chapman and Hall/CRC, 2015. A.5.1

[111] Sam Evens, Jiang-Hua Lu, and Alan Weinstein. Transverse measures, the modular class, and a cohomology pairing for Lie algebroids. *arXiv preprint dg-ga/9610008*, 1996. 4.3.1.6

[112] Gregory L Eyink, Joel L Lebowitz, and Herbert Spohn. Hydrodynamics and fluctuations outside of local equilibrium: driven diffusive systems. *Journal of Statistical physics*, 83(3-4):385–472, 1996. 4.6.3

[113] Gregory S Ezra. On the statistical mechanics of non-Hamiltonian systems: the generalized Liouville equation, entropy, and time-dependent metrics. *Journal of mathematical chemistry*, 35(1):29–53, 2004. 3.2.1

[114] Youhan Fang, Jesus-Maria Sanz-Serna, and Robert D Skeel. Compressible generalized hybrid Monte Carlo. *The Journal of chemical physics*, 140(17):174108, 2014. 1.1, 3.2.1, 3.4.2.3, 4, 4.3.3, 4.3.3, 4.4.5

[115] Matteo Fasiolo, Natalya Pya, Simon N Wood, et al. A comparison of inferential methods for highly nonlinear state space models in ecology and epidemiology. *Statistical Science*, 31(1):96–118, 2016. A.8

[116] Marián Fecko. *Differential geometry and Lie groups for physicists*. Cambridge University Press, 2006. 3.5.4.2

[117] Kang Feng and Mengzhao Qin. *Symplectic geometric algorithms for Hamiltonian systems*. Springer, 2010. 4.3.2

[118] Mario Fernández-Pendás, Elena Akhmatskaya, and JM Sanz-Serna. Adaptive multistage integrators for optimal energy conservation in molecular simulations. *Journal of Computational Physics*, 327:434–449, 2016. 2.1.3

[119] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-Ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. *arXiv preprint arXiv:1810.08278*, 2018. A.4.8

[120] Adriaan Daniël Fokker. Die mittlere Energie rotierender elektrischer Dipole im Strahlungsfeld. *Annalen der Physik*, 348(5):810–820, 1914. 4.6.3

[121] Gerald B Folland. *Real analysis: modern techniques and their applications*. John Wiley & Sons, 2013. A.2

[122] Theodore Frankel. *The geometry of physics: an introduction*. Cambridge university press, 2011. 5.2

[123] Kenji Fukumizu. Exponential manifold by reproducing kernel Hilbert spaces. *Algebraic and Geometric mothods in statistics*, pages 291–306, 2009. 2.2.1

[124] Daniel Gabay. Minimizing a differentiable function over a differential manifold. *Journal of Optimization Theory and Applications*, 37(2):177–219, 1982. 6

[125] Jean Gallier and Jocelyn Quaintance. Notes on differential geometry and Lie groups. *University of Pennsylvannia*, 4, 2012. 3.1.2, 3.1.3.1, 3.8.1.2, 3.8.2, 3.9.2

[126] Sylvestre Gallot, Dominique Hulin, and Jacques Lafontaine. *Riemannian geometry*, volume 3. Springer, 1990. 3.7.2

[127] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. *arXiv preprint arXiv:1706.00292*, 2017. A.4.6

[128] C. J. Geyer. On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):261–274, 1994. 6

[129] Charles J Geyer. Markov Chain Monte Carlo maximum likelihood. 1991. 6

[130] Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002. 5.4.1

[131] M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(2):123–214, 2011. 1, 14, 4.4.3, 5, 5.1.4, 5.2, A.5.5, A.5.5.1

[132] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. A.4.6

[133] J. Gorham, A. Duncan, L. Mackey, and S. Vollmer. Measuring sample quality with diffusions. *arXiv:1506.03039. To appear in Annals of Applied Probability.*, 2016. 6.1, 6.2

[134] J. Gorham and L. Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems*, pages 226–234, 2015. 6.1, 6.2

[135] J. Gorham and L. Mackey. Measuring sample quality with kernels. In *Proceedings of the International Conference on Machine Learning*, pages 1292–1301, 2017. 6.1, 6.3, 6.3

[136] Jackson Gorham, Andrew B Duncan, Sebastian J Vollmer, Lester Mackey, et al. Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5):2884–2928, 2019. 5.4.2, 6.9

[137] Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems*, pages 226–234, 2015. 5.4.2, 6.9

[138] Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1292–1301. JMLR. org, 2017. 5.5

[139] Matthew M Graham, Alexandre H Thiery, and Alexandros Beskos. Manifold Markov Chain Monte Carlo methods for Bayesian inference in a wide class of diffusion models. *arXiv preprint arXiv:1912.02982*, 2019. 3.5.2.1

[140] Robert Graham. Covariant formulation of non-equilibrium statistical thermodynamics. *Zeitschrift für Physik B Condensed Matter*, 26(4):397–405, 1977. 4, 4.6.3

[141] Clara Grazian and Yanan Fan. A review of approximate Bayesian computation methods via density estimation: Inference for simulator-models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2019. A.8

[142] Arthur Gretton. Introduction to RKHS, and some simple kernel algorithms. *Adv. Top. Mach. Learn. Lecture Conducted from University College London*, 16, 2013. 5.3

[143] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012. 5.5

[144] Rajan Gupta. Introduction to lattice QCD: Course. Technical report, 1997. A.4.1

[145] M. U. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010. 6

[146] M. U. Gutmann and A. Hyvarinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13:307–361, 2012. 6, 6

[147] E. Hairer, C. Lubich, and G. Wanner. *Geometric numerical integration algorithms for ordinary differential equations*. Springer, 2006. A.5.2

[148] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*, volume 31. Springer Science & Business Media, 2006. 3.5.2.1, 3.5.4.1, 11, 9

[149] Petr Hájek and Michal Johanis. *Smooth analysis in Banach spaces*, volume 19. Walter de Gruyter GmbH & Co KG, 2014. A.8.2

[150] Mark JD Hamilton. *Mathematical gauge theory*. Springer, 2017. A.4.1

[151] Carsten Hartmann. An ergodic sampling scheme for constrained Hamiltonian systems with applications to molecular dynamics. *Journal of Statistical Physics*, 130(4):687–711, 2008. 3.5.2.1

[152] Martin Hasenbusch. Speeding up the hybrid Monte Carlo algorithm for dynamical fermions. *Physics Letters B*, 519(1-2):177–182, 2001. A.4.1

[153] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002. 6, 6

[154] Liam Hodgkinson, Robert Salomone, and Fred Roosta. The reproducing Stein kernel approach for post-hoc corrected sampling. *arXiv preprint arXiv:2001.09266*, 2020. 6.9, 6.9, A.7.1

[155] Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, pages 293–325, 1948. 6.13.1.2

[156] Wassily Hoeffding. The strong law of large numbers for U-statistics. Technical report, North Carolina State University Department of Statistics, 1961. 6.13.3.1

[157] Christian Hoelbling. Lattice QCD: concepts, techniques and some results. *arXiv preprint arXiv:1410.3403*, 2014. A.4.1

[158] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013. 5.2

[159] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008. 5.3

[160] Andrew Holbrook. Note on the geodesic Monte Carlo. *arXiv preprint arXiv:1805.05289*, 2018. 1.1, 3, 3.2, 3.2.4, 3.6, 3.6.2, 3.6.4, 4.5.2.1

[161] Andrew Holbrook, Shiwei Lan, Alexander Vandenberg-Rodes, and Babak Shahbaba. Geodesic Lagrangian Monte Carlo over the space of positive definite matrices: with application to Bayesian spectral density estimation. *Journal of Statistical Computation and Simulation*, 2017. 1.1, 3, 3.2.4, 3.5, 3.6, 3.6.5, 3.8

[162] Andrew Holbrook, Alexander Vandenberg-Rodes, and Babak Shahbaba. Bayesian inference on matrix manifolds for linear dimensionality reduction. *arXiv preprint arXiv:1606.04478*, 2016. 1.1, 3.1.3.1.1, 3.7.1, 3.8

[163] Darryl D Holm. *Geometric mechanics: Dynamics and symmetry*, volume 1. Imperial College Press, 2008. 4, 3.9.1.2

[164] Darryl D. Holm. *Geometric Mechanics*. Imperial College Press, 2011. 3.2.2

[165] Darryl D Holm, Jerrold E Marsden, and Tudor S Ratiu. Euler-Poincaré models of ideal fluids with nonlinear dispersion. *Physical Review Letters*, 80(19):4173, 1998. 4.2.2

[166] Darryl D Holm, Tanya Schmah, and Cristina Stoica. *Geometric mechanics and symmetry: from finite to infinite dimensions*, volume 12 of *Oxford texts in applied and engineering mathematics*. Oxford University Press, 2009. 5, 3.8.3.3

[167] Lars Hörmander et al. Hypoelliptic second order differential equations. *Acta Mathematica*, 119:147–171, 1967. 18

[168] Alan M Horowitz. A generalized guided Monte Carlo algorithm. *Physics Letters B*, 268(2):247–252, 1991. 3.4.2.2

[169] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. Wiley, 2009. 6.7

[170] Jonathan Huggins and Lester Mackey. Random feature Stein discrepancies. In *Advances in Neural Information Processing Systems*, pages 1899–1909, 2018. 6.4

[171] Chii-Ruey Hwang, Shu-Yin Hwang-Ma, and Shuenn-Jyi Sheu. Accelerating Gaussian diffusions. *The Annals of Applied Probability*, pages 897–913, 1993. 4, 4.6.3

[172] Chii-Ruey Hwang, Shu-Yin Hwang-Ma, Shuenn-Jyi Sheu, et al. Accelerating diffusions. *The Annals of Applied Probability*, 15(2):1433–1444, 2005. 4.4.3

[173] A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–708, 2006. 6.1

[174] A. Hyvärinen. Some extensions of score matching. *Computational Statistics and Data Analysis*, 51(5):2499–2512, 2007. 6.1

[175] Aapo Hyvärinen. Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation. *Neural computation*, 11(7):1739–1768, 1999. 6.1

[176] Nobuyuki Ikeda and Shinzo Watanabe. *Stochastic differential equations and diffusion processes*. Elsevier, 2014. 4.4.3, 4.6.3, 6.9, A.6.1

[177] Jesús A Izaguirre and Scott S Hampton. Shadow hybrid Monte Carlo: an efficient propagator in phase space of macromolecules. *Journal of Computational Physics*, 200(2):581–604, 2004. A.4.2

[178] W. Jitkrittum, W. Xu, Z. Szabo, K. Fukumizu, and A. Gretton. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, pages 261–270, 2017. 6.4

[179] Balint Joo, Brian Pendleton, Anthony D Kennedy, Alan C Irving, James C Sexton, Stephen M Pickles, Stephen P Booth, UKQCD Collaboration, et al. Instability in the molecular dynamics step of a hybrid Monte Carlo algorithm in dynamical fermion lattice qcd simulations. *Physical Review D*, 62(11):114501, 2000. A.4.1

[180] Jürgen Jost and Jèurgen Jost. *Riemannian geometry and geometric analysis*, volume 42005. Springer, 2008. 3.9.2.4

[181] Sham M Kakade. A natural policy gradient. In *Advances in neural information processing systems*, pages 1531–1538, 2002. 5.2

[182] Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018. 5.3

[183] R. Karakida, M. Okada, and S.-I. Amari. Adaptive natural gradient learning algorithms for unnormalized statistical models. *Artificial Neural Networks and Machine Learning - ICANN*, 2016. 6.6

[184] Ryo Karakida, Masato Okada, and Shun-ichi Amari. Adaptive natural gradient learning algorithms for unnormalized statistical models. In *International Conference on Artificial Neural Networks*, pages 427–434. Springer, 2016. 5.2

[185] Mehran Kardar. *Statistical physics of particles*. Cambridge University Press, 2007. 3.2.2.1

[186] A. D. Kennedy and Pietro Rossi. Classical Mechanics on Group Manifolds. *Nucl. Phys.*, B327:782–790, 1989. 1.1, 2.1, 3, 3.8, 3.9, 3.9.1.1, 3.9.1.1, 3.9.1.2, A.5.7.3

[187] A. D. Kennedy, P. J. Silva, and M. A. Clark. Shadow Hamiltonians, Poisson Brackets, and Gauge Theories. *Physical Review*, D87(3):034511, 2013. 1.1, 2.1.3.1, 3.5.4.2, 3.9, 3.9.1.2, 3.9.1.2, 3.9.2.1, A.5.7.3, A.5.7.6

[188] AD Kennedy and Brian Pendleton. Cost of the generalised hybrid Monte Carlo algorithm for free field theory. *Nuclear Physics B*, 607(3):456–510, 2001. 3.4.2.2

[189] Anthony D Kennedy, Michael A Clark, and PJ Silva. Force gradient integrators. *arXiv preprint arXiv:0910.2950*, 2009. A.4.1

[190] Achim Klenke. *Probability theory: a comprehensive course*. Springer Science & Business Media, 2013. A.2

[191] Anthony W Knapp. *Lie groups beyond an introduction*, volume 140. Springer Science & Business Media, 2013. 3.1.1

[192] Francesco Knechtli, Michael Günther, and Michael Peardon. *Lattice quantum chromodynamics: practical essentials*. Springer, 2017. 3.9, 3.9.1.2

[193] Andrei Kolmogoroff. Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Mathematische Annalen*, 104(1):415–458, 1931. 4.6.3

[194] Tomasz Komorowski, Claudio Landim, and Stefano Olla. *Fluctuations in Markov processes: time symmetry and martingale approximation*, volume 345. Springer Science & Business Media, 2012. 4.4.2

[195] Yvette Kosmann-Schwarzbach. Poisson manifolds, Lie algebroids, modular classes: a survey. *Symmetry, Integrability and Geometry: Methods and Applications*, 2008. 4.3.1.6

[196] Urs Köster and Aapo Hyvärinen. A two-layer model of natural stimuli estimated with score matching. *Neural Computation*, 22(9):2308–2333, 2010. 6

[197] Jean-Louis Koszul. Crochet de Schouten-Nijenhuis et cohomologie. *Astérisque*, 137:257–271, 1985. 4.2.1, 4.3.1.3

[198] S. Kotz, T. J. Kozubowski, and K. Podgorski. *The Laplace Distribution and Generalizations*. Springer, 2001. 6.8.1

[199] Oldřich Kowalski. Riemannian manifolds with homogeneous geodesics. In *Advances in Differential Geometry and Topology*, pages 47–57. World Scientific, 1990. 3.1.3.2

[200] Yury A Kutoyants. *Statistical inference for ergodic diffusion processes*. Springer Science & Business Media, 2013. 2.2

[201] Jeroen SW Lamb and John AG Roberts. Time-reversal symmetry in dynamical systems: a survey. *Physica-Section D*, 112(1):1–39, 1998. 4.3.3

[202] Shiwei Lan, Vasileios Stathopoulos, Babak Shahbaba, and Mark Girolami. Markov Chain Monte Carlo from Lagrangian dynamics. *Journal of Computational and Graphical Statistics*, 24(2):357–378, 2015. 1.1, 3.1, 3.2, 3.2.4, 3.5

[203] Shiwei Lan, Jeffrey Streets, and Babak Shahbaba. Wormhole Hamiltonian Monte Carlo. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014. 2.1.2

[204] Shiwei Lan, Bo Zhou, and Babak Shahbaba. Spherical Hamiltonian Monte Carlo for Constrained Target Distributions. In Eric P. Xing and Tony Jebara, editors, *31st International Conference on Machine Learning, Beijing, China*, volume 32 of *JMLR workshop and conference proceedings*, pages 629–637, 2014. 3.2.4, 3.8

[205] LD Landau and EM Lifshitz. *Course of theoretical physics. vol. 1: Mechanics*. Oxford, 1960. 4.2.2

[206] Longin Jan Latecki, Aleksandar Lazarevic, and Dragoljub Pokrajac. Outlier detection with kernel density functions. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 61–75. Springer, 2007. 2.2

[207] Diego S Ledesma. Lyapunov 1-form for diffusion processes. *arXiv preprint arXiv:1205.5582*, 2012. 3

[208] John M Lee. *Riemannian manifolds: an introduction to curvature*, volume 176. Springer Science & Business Media, 2006. 3.6.2.1, 3.6.2.2

[209] John M Lee. Smooth manifolds. In *Introduction to Smooth Manifolds*, pages 1–31. Springer, 2013. 3.5.2, 4.2, 6.9

[210] B. Leimkuhler and G. W. Patrick. A symplectic integrator for Riemannian Manifolds. *Journal of Nonlinear Science*, 1996. 3.5.1, 4.5.2.1

[211] Benedict Leimkuhler and Charles Matthews. Efficient molecular dynamics using geodesic integration and solvent–solute splitting. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 472(2189):20160138, 2016. 3.1.3, 3.5.2.1

[212] Benedict Leimkuhler and Sebastian Reich. *Simulating Hamiltonian dynamics*, volume 14. Cambridge university press, 2004. 3.2.2, 6, 3.5.1, 3.5.2.1, 3.5.4.3

[213] Tony Lelièvre, Francis Nier, and Grigorios A Pavliotis. Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion. *Journal of Statistical Physics*, 152(2):237–274, 2013. 4.4.3, 4.5.1

[214] Tony Lelièvre, Mathias Rousset, and Gabriel Stoltz. Hybrid Monte Carlo methods for sampling probability measures on submanifolds. *arXiv preprint arXiv:1807.02356*, 2018. 3.5.2.1

[215] Tony Lelievre and Gabriel Stoltz. Partial differential equations and stochastic methods in molecular dynamics. *Acta Numerica*, 25:681–880, 2016. 4.5

[216] Melvin Leok and Jun Zhang. Connecting information geometry and geometric mechanics. *Entropy*, 19(10):518, 2017. A.5.6.5

[217] Christophe Ley, Gesine Reinert, Yvik Swan, et al. Distances between nested densities and a measure of the impact of the prior in Bayesian statistics. *The Annals of Applied Probability*, 27(1):216–241, 2017. 5.4.1

[218] Christophe Ley, Yvik Swan, et al. Stein's density approach and information inequalities. *Electronic Communications in Probability*, 18, 2013. 5.4.1, 6.9

[219] Y. Li and R. E. Turner. Gradient estimators for implicit models. In *International Conference on Learning Representations*, 2018. 6.2

[220] Paulette Libermann. On Liouville forms. *Banach Center Publications*, 51(1):151–164, 2000. 3.3

[221] Paulette Libermann and Charles-Michel Marle. *Symplectic geometry and analytical mechanics*, volume 35. Springer Science & Business Media, 2012. 3.3

[222] Sophus Lie. Vorlesungen über Differentialgleichungen mit bekannten infinitesimalen Transformationen. *Bull. Amer. Math. Soc*, 4:155–167, 1898. 4.6.4

[223] Chang Liu and Jun Zhu. Riemannian Stein Variational Gradient Descent for Bayesian Inference. (i), 2017. 6.3

[224] Chang Liu and Jun Zhu. Riemannian Stein variational gradient descent for Bayesian inference. In *Thirty-second aaai conference on artificial intelligence*, 2018. 6.9, 6.9

[225] Chang Liu, Jun Zhu, and Yang Song. Stochastic gradient geodesic MCMC methods. In *Advances in neural information processing systems*, pages 3009–3017, 2016. 4.5.2.1

[226] Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of the International Conference on Machine Learning*, pages 276–284, 2016. 6.3, 6.3, 6.3, 6.4

[227] Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, 2016. 6.3

[228] Q. Liu and D. Wang. Learning deep energy models: Contrastive divergence vs. amortized mle. *arXiv preprint arXiv:1707.00797*, 2017. 6, 6.3

[229] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284, 2016. 5.5, 6.3

[230] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in neural information processing systems*, pages 2378–2386, 2016. 5.2

[231] S. Liu, T. Kanamori, W. Jitkrittum, and Y. Chen. Fisher efficient inference of intractable models. *arXiv:1805.07454*, 2018. 6.2, 6.8.4

[232] Yi Liu and Mark E Tuckerman. Generalized Gaussian moment thermostatting: A new continuous dynamical approach to the canonical ensemble. *The Journal of Chemical Physics*, 112(4):1685–1700, 2000. 3.2.1

[233] Zhang-Ju Liu and Ping Xu. On quadratic Poisson structures. *letters in mathematical physics*, 26(1):33–42, 1992. 4.3.1.6

[234] S. Livingstone and M. Girolami. Information-Geometric Markov Chain Monte Carlo Methods Using Diffusions. *Entropy*, 16(6):3074–3102, 2014. 4.4.3

[235] Samuel Livingstone, Michael Betancourt, Simon Byrne, and Mark Girolami. On the geometric ergodicity of Hamiltonian Monte Carlo. *arXiv preprint arXiv:1601.08057*, 2016. A.4.5

[236] Samuel Livingstone, Michael F Faulkner, and Gareth O Roberts. Kinetic energy choice in Hamiltonian/hybrid Monte Carlo. *Biometrika*, 106(2):303–319, 2019. 2.1.2

[237] Samuel Livingstone and Mark Girolami. Information-geometric Markov Chain Monte Carlo methods using diffusions. *Entropy*, 16(6):3074–3102, 2014. 5.1.4

[238] Jean-Louis Loday and Teimuraz Pirashvili. Universal enveloping algebras of Leibniz algebras and (co) homology. *Mathematische Annalen*, 296(1):139–158, 1993. 4.3.1.5

[239] S. Lyu. Interpretation and generalization of score matching. In *Conference on Uncertainty in Artificial Intelligence*, pages 359–366, 2009. 6.4

[240] Siwei Lyu. Interpretation and generalization of score matching. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 359–366. AUAI Press, 2009. 6.1, 6.10

[241] C. Ma and D. Barber. Black-box Stein divergence minimization for learning latent variable models. *Advances in Approximate Bayesian Inference, NIPS 2017 Workshop*, 2017. 6.2

[242] Yi-An Ma, Tianqi Chen, and Emily Fox. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, pages 2917–2925, 2015. 1.1, 4, 4.1, 4.1, 4.4, 4.4.2, 4.4.4, 22, 4.4.4, 16, 4.6.3, 7

[243] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003. 2.1

[244] Paul B Mackenzie. An improved hybrid Monte Carlo method. *Physics Letters B*, 226(3-4):369–371, 1989. A.4.5

[245] L. Mackey and J. Gorham. Multivariate Stein factors for a class of strongly log-concave distributions. *Electronic Communications in Probability*, 21, 2016. 6.1

[246] Kanti V Mardia, John T Kent, and Arnab K Laha. Score matching estimators for directional distributions. *arXiv preprint arXiv:1604.08470*, 2016. 6.1

[247] Jean-Michel Marin, Pierre Pudlo, Christian P Robert, and Robin J Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012. A.8

[248] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov Chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003. A.8

[249] Charles-Michel Marle et al. On Henri Poincaré's note "sur une forme nouvelle des équations de la mécanique". *Journal of geometry and symmetry in physics*, 29:1–38, 2013. 3.2.4.1

[250] Jerrold E. Marsden. *Lectures on Mechanics*, volume 174 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, 1992. October, 2009. 3.5, 3.8.1.1, 3.8.1.2, 3.8.2

[251] Jerrold E Marsden and Tudor S Ratiu. *Introduction to mechanics and symmetry: a basic exposition of classical mechanical systems*, volume 17. Springer Science & Business Media, 2013. 3.3.1, 3.8.1.3, 4.2.2.2, A.5.3, A.5.4

[252] Jerrold E Marsden and Matthew West. Discrete mechanics and variational integrators. *Acta Numerica*, 10:357–514, 2001. 3.5.1, 3.5.2, A.5.3, A.5.5, A.5.5, A.5.5.2, A.5.5.3

[253] Robert Martin and Patrizio Neff. Minimal geodesics on GL(n) for left-invariant, right-O(n)-invariant Riemannian metrics. *arXiv preprint arXiv:1409.7849*, 2014. 3.6.3, 3.9.2

[254] Glenn J Martyna, Michael L Klein, and Mark Tuckerman. Nosé–Hoover chains: The canonical ensemble via continuous dynamics. *The Journal of chemical physics*, 97(4):2635–2643, 1992. 3.2.1

[255] Youssef Marzouk, Tarek Moselhy, Matthew Parno, and Alessio Spantini. An introduction to sampling via measure transport. *arXiv preprint arXiv:1602.05023*, 2016. 3.2.2, 3.6.1

[256] J Masoliver, L Garrido, and J Llosa. Geometrical derivation of the intrinsic Fokker-Planck equation and its stationary distribution. *Journal of statistical physics*, 46(1-2):233–248, 1987. 4.6.3

[257] R McLachlan and Matthew Perlmutter. Integrators for nonholonomic mechanical systems. *Journal of nonlinear science*, 16(4):283–328, 2006. A.5.5

[258] Robert I McLachlan. On the numerical integration of ordinary differential equations by symmetric composition methods. *SIAM Journal on Scientific Computing*, 16(1):151–168, 1995. 2.1.3

[259] Robert I McLachlan, Klas Modin, Olivier Verdier, and Matt Wilkins. Geometric generalisations of SHAKE and RATTLE. *Foundations of Computational Mathematics*, 14(2):339–370, 2014. 3.5.2.1, 7

[260] Robert I McLachlan and G Reinout W Quispel. Splitting methods. *Acta Numerica*, 11:341–434, 2002. 4.2.1, 4.4.2

[261] Charles A Micchelli and Massimiliano Pontil. Kernels for multi–task learning. In *Advances in neural information processing systems*, pages 921–928, 2005. 5.3

[262] Charles A Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural computation*, 17(1):177–204, 2005. 5.3, 6.11.1

[263] Mario Micheli and Joan Alexis Glaunes. Matrix-valued kernels for shape deformation analysis. *arXiv preprint arXiv:1308.5739*, 2013. 5.3, 6.11

[264] Peter W Michor. *Topics in differential geometry*, volume 93. American Mathematical Soc., 2008. 3.7, 4.2

[265] Alexander Mielke. Finite elastoplasticity Lie groups and geodesics on sl (d). In *Geometry, mechanics, and dynamics*, pages 61–90. Springer, 2002. 3.6.3, 3.9.2

[266] Nina Miolane and Xavier Pennec. Computing bi-invariant pseudo-metrics on Lie groups for consistent statistics. *Entropy*, 17(4):1850–1881, 2015. 3.9.2.3

[267] A. Mnih and Y. W. Teh. A fast and simple algorithm for training neural probabilistic language models. In *Proceedings of the International Conference on Machine Learning*, pages 419–426, 2012. 6

[268] Maher Moakher and Mourad Zéraï. The Riemannian geometry of the space of positive-definite matrices and its application to the regularization of positive-definite matrix-valued data. *Journal of Mathematical Imaging and Vision*, 40(2):171–187, 2011. 3.6.5

[269] K Modin, M Perlmutter, S Marsland, and R McLachlan. Geodesics on Lie groups: Euler equations and totally geodesic subgroup. 2010. 3.9.3.1

[270] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016. A.4.6

[271] Philip J Morrison. Hamiltonian description of the ideal fluid. *Reviews of modern physics*, 70(2):467, 1998. 4.2.2, 4.3.1.5

[272] Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev GAN. *arXiv preprint arXiv:1711.04894*, 2017. A.4.6

[273] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017. 5.5, 5.5

[274] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997. 5.4.1

[275] Radford M Neal. Improving asymptotic variance of MCMC estimators: Non-reversible chains are better. *arXiv preprint math/0407281*, 2004. 3.4.2.3

[276] Radford M. Neal. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, chapter 5. Chapman & Hall/CRC Press, 2011. 3.4.1

[277] W. K. Newey and D. McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, 1994. 2.2, 6.13.1.1, 6.13.1.2, 6.13.2.1, 6.13.2.2

[278] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010. A.4.6

[279] James R Norris. A complete differential formalism for stochastic calculus in manifolds. *Séminaire de probabilités de Strasbourg*, 26:189–209, 1992. 4.4.1

[280] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pages 271–279, 2016. A.4.6

[281] C. J. Oates, J. Cockayne, F.-X. Briol, and M. Girolami. Convergence Rates for a Class of Estimators Based on Stein's Identity. *arXiv preprint arXiv:1603.03220*, 2016. 6.9

[282] C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society B: Statistical Methodology*, 79(3):695–718, 2017. 6.3, 6.3, 6.9

[283] Chris J Oates, Mark Girolami, and Nicolas Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718, 2017. 5.5

[284] IP Omelyan, IM Mryglod, and R Folk. Symplectic analytically integrable decomposition algorithms: classification, derivation, and application to molecular dynamics, quantum and celestial mechanics simulations. *Computer Physics Communications*, 151(3):272–314, 2003. 3.5.4.2, A.4.1

[285] Barrett O'neill. *Semi-Riemannian geometry with applications to relativity*, volume 103. Academic press, 1983. 3.6.2.1

[286] Juan-Pablo Ortega and Tudor S Ratiu. Cotangent bundle reduction. *arXiv preprint math/0508636*, 2005. 3.8.1.2, 3.8.1.3

[287] Hirofumi Osada et al. Diffusion processes with generators of generalized divergence form. *Journal of Mathematics of Kyoto University*, 27(4):597–619, 1987. 4.6.3

[288] Michela Ottobre. Markov Chain Monte Carlo and irreversibility. *Reports on Mathematical Physics*, 77(3):267–292, 2016. 1.1, 3.4.2.3

[289] Michela Ottobre, Natesh S Pillai, Frank J Pinski, Andrew M Stuart, et al. A function space HMC algorithm with second order Langevin diffusion limit. *Bernoulli*, 22(1):60–106, 2016. 3.4.2.3, 4.5.2, 4.5.2.2

[290] Hyeyoung Park, S-I Amari, and Kenji Fukumizu. Adaptive natural gradient learning algorithms for various stochastic models. *Neural Networks*, 13(7):755–764, 2000. 5.2

[291] Matthew Parry, A Philip Dawid, Steffen Lauritzen, et al. Proper local scoring rules. *The Annals of Statistics*, 40(1):561–592, 2012. 2.2.1

[292] Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962. 2.2

[293] Vern I Paulsen and Mrinal Raghupathi. *An introduction to the theory of reproducing kernel Hilbert spaces*, volume 152. Cambridge University Press, 2016. 5.3

[294] Grigorios A Pavliotis. *Stochastic processes and applications: diffusion processes, the Fokker-Planck and Langevin equations*, volume 60. Springer, 2014. 4.1, 4.1, 4.4.1, 4.4.1, 4.4.5, 6, 6.5, A.4.10

[295] Xavier Pennec. *Probabilities and statistics on Riemannian manifolds: A geometric approach*. PhD thesis, INRIA, 2004. 3.1.2, A.5.1

[296] Xavier Pennec. Intrinsic statistics on Riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127, 2006. 3.1

[297] Xavier Pennec and Nicholas Ayache. Uniform distribution, distance and expectation problems for geometric features processing. *Journal of Mathematical Imaging and Vision*, 9(1):49–67, 1998. 3.1.2

[298] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 2.2.1, A.4.8

[299] S. Pigola and A. G. Setti. Global divergence theorems in nonlinear PDEs and geometry. *Ensaios Matemáticos*, 26:1–77, 2014. 2, 6.11.3.1

[300] VM Planck. Über einen Satz der statistischen Dynamik und seine Erweiterung in der Quantentheorie. *Sitzungberichte der*, 1917. 4.6.3

[301] Cristian Predescu, Ross A Lippert, Michael P Eastwood, Douglas Ierardi, Huafeng Xu, Morten O Jensen, Kevin J Bowers, Justin Gullingsrud, Charles A Rendleman, Ron O Dror, et al. Computationally efficient molecular dynamics integrators with improved sampling accuracy. *Molecular Physics*, 110(9-10):967–983, 2012. 2.1.3

[302] Tijana Radivojević and Elena Akhmatskaya. Modified Hamiltonian Monte Carlo for Bayesian inference. *Statistics and Computing*, pages 1–28, 2019. 3.4.2.3, A.4.2

[303] Tijana Radivojević, Mario Fernández-Pendás, Jesús María Sanz-Serna, and Elena Akhmatskaya. Multi-stage splitting integrators for sampling with modified Hamiltonian Monte Carlo methods. *Journal of Computational Physics*, 373:900–916, 2018. 3.5.4.2

[304] R. Ranganath, J. Altosaar, D. Tran, and D. M. Blei. Operator variational inference. In *Advances in Neural Information Processing Systems*, pages 496–504, 2016. 6.2

[305] Luc Rey-Bellet and Konstantinos Spiliopoulos. Irreversible langevin samplers and variance reduction: a large deviations approach. *Nonlinearity*, 28(7):2081, 2015. 4.4.3

[306] Hannes Risken. Fokker-Planck equation. In *The Fokker-Planck Equation*, pages 63–95. Springer, 1996. 4.6.3

[307] Gareth O Roberts, Richard L Tweedie, et al. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. 4.1.2

[308] Miguel Rodríguez-Olmos. The isotropy lattice of a lifted action. *Comptes Rendus Mathematique*, 343(1):41–46, 2006. 18

[309] Nathan Ross et al. Fundamentals of Stein's method. *Probability Surveys*, 8:210–293, 2011. 5.4.1

[310] S. Roth and M. J. Black. Fields of experts. *International Journal of Computer Vision*, 82(2):205, 2009. 6

[311] Christof Seiler, Simon Rubinstein-Salzedo, and Susan Holmes. Positive curvature and Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 586–594, 2014. A.5.4

[312] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, Kenji Fukumizu, et al. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013. 5.5

[313] Alessandro Sergi and Mauro Ferrario. Non-Hamiltonian equations of motion with a conserved energy. *Physical Review E*, 64(5):056125, 2001. 3.2.1

[314] Alessandro Sergi and Paolo V Giaquinta. On the geometry and entropy of non-Hamiltonian phase space. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(02):P02013, 2007. 3.2.1

[315] JC Sexton and DH Weingarten. Hamiltonian evolution for the hybrid Monte Carlo algorithm. *Nuclear Physics B*, 380(3):665–677, 1992. 2.1.3.1, A.4.1

[316] Babak Shahbaba, Shiwei Lan, Wesley O Johnson, and Radford M Neal. Split Hamiltonian Monte Carlo. *Statistics and Computing*, 24(3):339–349, 2014. 2.1.3.1

[317] Babak Shahbaba, Luis Martinez Lomeli, Tian Chen, and Shiwei Lan. Deep Markov Chain Monte Carlo. *arXiv preprint arXiv:1910.05692*, 2019. 3.8.1.2, A.4.7

[318] Jonathan Z Simon. Higher-derivative Lagrangians, nonlocality, problems, and solutions. *Physical Review D*, 41(12):3720, 1990. A.4.2

[319] Robert D Skeel and David J Hardy. Practical construction of modified Hamiltonians. *SIAM Journal on Scientific Computing*, 23(4):1172–1188, 2001. A.4.2

[320] J. Sohl-dickstein, P. Battaglino, and M. R. DeWeese. Minimum probability flow learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 905–912, 2011. 6, 6.5

[321] Jascha Sohl-Dickstein, Mayur Mudigonda, and Michael DeWeese. Hamiltonian Monte Carlo without detailed balance. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 719–726, Bejing, China, 22–24 Jun 2014. PMLR. 3.4.2.2

[322] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *arXiv preprint arXiv:1907.05600*, 2019. A.4.7

[323] Luis Antonio Santaló Sors and Luis A Santaló. *Integral geometry and geometric probability*. Cambridge university press, 2004. 3.1.2

[324] B. Sriperumbudur, K. Fukumizu, A. Gretton, A. Hyvärinen, and R. Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(1):1830–1888, 2017. 6

[325] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010. 6.3

[326] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010. 5.5, 5.5, 6.3

[327] Charles Stein et al. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California, 1972. 5.4.1

[328] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008. 5.5, A.8.2

[329] Gabriel Stoltz, Mathias Rousset, et al. *Free energy computations: A mathematical perspective*. World Scientific, 2010. 2.1, 2.1, 3.4.1, 7, 3.4.2.3, 4, 4.1.1, 4.1.2, 4.4.5, A.4.5

[330] Heiko Strathmann, Dino Sejdinovic, Samuel Livingstone, Zoltan Szabo, and Arthur Gretton. Gradient-free Hamiltonian Monte Carlo with efficient kernel exponential families. In *Advances in Neural Information Processing Systems*, pages 955–963, 2015. A.4.3, A.4.7

[331] Christopher R Sweet, Scott S Hampton, Robert D Skeel, and Jesús A Izaguirre. A separable shadow Hamiltonian hybrid Monte Carlo method. *The Journal of chemical physics*, 131(17):174106, 2009. A.4.2

[332] Tetsuya Takaishi and Philippe De Forcrand. Odd-flavor hybrid Monte Carlo algorithm for lattice QCD. *International Journal of Modern Physics C*, 13(03):343–365, 2002. A.4.1

[333] Tetsuya Takaishi and Philippe De Forcrand. Testing and tuning symplectic integrators for the hybrid Monte Carlo algorithm in lattice QCD. *Physical Review E*, 73(3):036706, 2006. A.4.1

[334] Asuka Takatsu et al. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4):1005–1026, 2011. 5.1.2

[335] Molei Tao. Explicit symplectic approximation of nonseparable Hamiltonians: Algorithm and long time performance. *Physical Review E*, 94(4):043303, 2016. 3.5.2.1, 3.8, 5.1.4

[336] Vasily E Tarasov. Phase-space metric for non-Hamiltonian systems. *Journal of Physics A: Mathematical and General*, 38(10):2145, 2005. 3.2.1

[337] Clifford Henry Taubes. *Differential geometry: Bundles, connections, metrics and curvature*, volume 23. OUP Oxford, 2011. 3.6.3

[338] Michael Eugene Taylor. *Partial differential equations. 1, Basic theory*. Springer, 1996. 4.6.4

[339] George R Terrell, David W Scott, et al. Variable kernel density estimation. *The Annals of Statistics*, 20(3):1236–1265, 1992. 2.2

[340] James Thompson. Approximation of Riemannian measures by Stein's method. *arXiv preprint arXiv:2001.09910*, 2020. 6.9

[341] Luke Tierney et al. A note on Metropolis-Hastings kernels for general state spaces. *The Annals of Applied Probability*, 8(1):1–9, 1998. 3.4.1

[342] Nilesh Tripuraneni, Mark Rowland, Zoubin Ghahramani, and Richard Turner. Magnetic Hamiltonian Monte Carlo. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3453–3461. JMLR. org, 2017. 3.8.1.2, 3.8.1.3, 4.6.4

[343] Mark Tuckerman. *Statistical mechanics: theory and molecular simulation*. Oxford university press, 2010. 3.2.1

[344] Mark E Tuckerman, Yi Liu, Giovanni Ciccotti, and Glenn J Martyna. Non-Hamiltonian molecular dynamics: Generalizing Hamiltonian phase space principles to non-Hamiltonian systems. *The Journal of Chemical Physics*, 115(4):1678–1702, 2001. 3.2.1

[345] MBBJM Tuckerman, Bruce J Berne, and Glenn J Martyna. Reversible multiple time scale molecular dynamics. *The Journal of chemical physics*, 97(3):1990–2001, 1992. 2.1.3, 2.1.3.1

[346] ME Tuckerman, CJ Mundy, and GJ Martyna. On the classical statistical mechanics of non-Hamiltonian systems. *EPL (Europhysics Letters)*, 45(2):149, 1999. 3.2.1

[347] Daniele Valtorta and Giona Veronelli. Stokes'theorem, volume growth and parabolicity. *Tohoku Mathematical Journal, Second Series*, 63(3):397–412, 2011. 2

[348] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000. 2.2, 2.2.1, A.2

[349] Wilfred F Van Gunsteren and Martin Karplus. Effect of constraints on the dynamics of macromolecules. *Macromolecules*, 15(6):1528–1544, 1982. 3.5.2

[350] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998. 2.2

[351] C. Villani. *Optimal transport. Old and new*. Springer, 2009. A.2

[352] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008. 2.2.1, 2.2.1

[353] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. A.4.7

[354] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008. A.4.7

[355] Vito Volterra et al. Sur une genéralisation de la théorie des fonctions d'une variable imaginaire: Ier Mémoire. *Acta Mathematica*, 12:233–286, 1889. 4.2.1, 4.6.3

[356] Jessika Walter, Carsten Hartmann, and John H Maddocks. Ambient space formulations and statistical mechanics of holonomically constrained Langevin systems. *The European Physical Journal Special Topics*, 200(1):153–181, 2011. 3.5.2.1, 3.8, 4

[357] Ziyu Wang, Shakir Mohamed, and Nando Freitas. Adaptive Hamiltonian and Riemann manifold Monte Carlo. In *International Conference on Machine Learning*, pages 1462–1470, 2013. A.4.5

[358] Alan Weinstein. The modular automorphism group of a Poisson manifold. *Journal of Geometry and Physics*, 23(3-4):379–394, 1997. 4, 4.3, 4.3.1, 4.3.1.6

[359] Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004. 5.3

[360] L. Wenliang, D. Sutherland, H. Strathmann, and A. Gretton. Learning deep kernels for exponential family densities. *arXiv:1811.08357*, 2018. 6

[361] Kenneth G Wilson. Confinement of quarks. *Physical review D*, 10(8):2445, 1974. A.4.1

[362] Simon N Wood. Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102, 2010. A.8

[363] Viktor Yanush and Dmitry Kropotov. Hamiltonian monte-carlo for orthogonal matrices. *arXiv preprint arXiv:1901.08045*, 2019. A.5.2

[364] I.-K. Yeo and R. A. Johnson. A uniform strong law of large numbers for U-statistics with application to transforming to near symmetry. *Statistics & Probability Letters*, 51(1):63–69, 2001. 6.13.1.1, 6.13.1.2

[365] Anna Yershova, Swati Jain, Steven M Lavalle, and Julie C Mitchell. Generating uniform incremental grids on SO (3) using the Hopf fibration. *The International journal of robotics research*, 29(7):801–812, 2010. 3.1.2

[366] Haruo Yoshida. Construction of higher order symplectic integrators. *Physics letters A*, 150(5-7):262–268, 1990. 2.1.3

[367] Emilio Zappa, Miranda Holmes-Cerfon, and Jonathan Goodman. Monte Carlo on manifolds: sampling densities and integrating functions. *Communications on Pure and Applied Mathematics*, 71(12):2609–2647, 2018. 3.5.2.1

[368] Cheng Zhang, Babak Shahbaba, and Hongkai Zhao. Hamiltonian Monte Carlo acceleration using surrogate functions with random bases. *Statistics and computing*, 27(6):1473–1490, 2017. A.4.3

[369] Cheng Zhang, Babak Shahbaba, and Hongkai Zhao. Precomputing strategy for Hamiltonian Monte Carlo method based on regularity in parameter space. *Computational Statistics*, 32(1):253–279, 2017. A.4.3

[370] Cheng Zhang, Babak Shahbaba, Hongkai Zhao, et al. Variational Hamiltonian Monte Carlo via score matching. *Bayesian Analysis*, 13(2):485–506, 2018. A.4.3

[371] Ge Zhong and Jerrold E Marsden. Lie-Poisson Hamilton-Jacobi theory and Lie-Poisson integrators. *Physics Letters A*, 133(3):134–139, 1988. 4.3.2

[372] Vladimir Mikhailovich Zolotarev. Probability metrics. *Teoriya Veroyatnostei i ee Primeneniya*, 28(2):264–287, 1983. A.2

Appendix A

Appendix

A.1 Vector Calculus

The following section contains background and important identities from vector calculus. For a function $g \in \Gamma(\mathcal{X}, \mathbb{R})$, $v \in \Gamma(\mathcal{X}, \mathbb{R}^d)$ and $A \in \Gamma(\mathcal{X}, \mathbb{R}^{d\times d})$ with components A_{ij}, v_i, g, we have $(\nabla g)_i = \partial_i g$, $(v \cdot A)_i = v_j A_{ji} = (v^\top A)_i$, $(\nabla \cdot A)_i = \partial_j A_{ji}$ which must be interpreted as the components of row-vectors; $(Av)_i = A_{ij} v_j$ which are the components of a column vector. Moreover $(\nabla v)_{ij} = \partial_j v_i$, $\nabla^2 f \equiv \nabla(\nabla f)$, $A : B \equiv \langle A, B\rangle = \mathrm{tr}(A^\top B) = A_{ij} B_{ij}$. We have the following identities (where in the last equality we treat $\nabla \cdot A$ and ∇g as column vectors)

$$
\begin{aligned}
\nabla \cdot (gv) &= \partial_i (g v_i) = v_i \partial_i g + g \partial_i v_i = (\nabla g) v + g \nabla \cdot v = \nabla g \cdot v + g \nabla \cdot v, \\
\nabla \cdot (gA) &= \partial_i (g A_{ij}) e_j = (A_{ij} \partial_i g + g \partial_i A_{ij})\, e_j = \nabla g \cdot A + g \nabla \cdot A = \nabla g^\top A + g \nabla \cdot A, \\
\nabla \cdot (Av) &= \partial_i (A_{ij} v_j) = (\nabla \cdot A) v + \mathrm{tr}[A \nabla v] = (\nabla \cdot A) \cdot v + \mathrm{tr}[A \nabla v].
\end{aligned}
$$

A.2 Convergence of Measures

Let $\mathcal{M}$ be a topological space (with measurable sets defined by the Borel sigma algebra). We are interested in sequence of measures (μ_n) converging to a measure μ in the sense that

$$
\int f \mathrm{d}\mu_n \to \int f \mathrm{d}\mu, \quad \forall f \in \mathcal{C}, \tag{A.1}
$$

for a given space of functions $\mathcal{C}$. When $\mathcal{C} = C_b(\mathcal{M})$ this is called the **narrow topology**. In particular when the measures are the distribution of random variables, $\mu_n \equiv P_{X_n}$, $\mu \equiv P_X$, we say (X_n) **converges in distribution** to X, if the laws (P_{X_n}) narrowly converge to P_X. See Lemma 18.9 [348] for equivalent ways to define convergence in distribution when $\mathcal{M}$ is a metric space. When $\mathcal{M}$ is locally compact and Hausdorff (LCH), and $\mathcal{R}(\mathcal{M})$ is the normed space of complex Radon measures with total variation norm, the Riesz representation theorem (7.17) [121] states the map $\mathcal{R}(\mathcal{M}) \to C_0^*(\mathcal{M})$ defined by $\mu \mapsto \left(f \mapsto \int f \mathrm{d}\mu\right)$ is an isometric isomorphism, where $C_0(\mathcal{M})$ is the Banach space of continuous functions on $\mathcal{M}$ that vanish at infinity with supremum norm (which is the closure of $C_c(\mathcal{M})$ by prop 4.35 [121]). Using this identification we may equip $\mathcal{R}(\mathcal{M})$ with the topology of weak* (i.e., pointwise) convergence of $C_0^*(\mathcal{M})$, that is $T_n \to T$ iff $T_n(f) \to T(f)$ for all $f \in C_0(\mathcal{M})$. Thus if $\mathcal{M}$ is LCH, the choice $\mathcal{C} \equiv C_0(\mathcal{M})$ above defines the **weak* topology**. In general the topology generated by $\mathcal{C} = C_c(\mathcal{M}), C_0(\mathcal{M}), C_b(\mathcal{M})$ are different unless $\mathcal{M}$ is compact. However by Portemanteau (theorem 13.16 [190]) if $\mathcal{M}$ is locally compact, Polish, and we restrict to probability meausres, then convergence in $C_c(\mathcal{M})$ is equivalent to convergence in $C_b(\mathcal{M})$. Hence for second countable LCH spaces (which are Polish), since $C_c(\mathcal{M}) \subset C_0(\mathcal{M}) \subset C_b(\mathcal{M})$ the three notions are equivalent on probability measures. Additional notions of convergence can be defined using **probability metrics**, which are metrics on the set of probability measures that are allowed to take the value $+\infty$ [372]. Many examples belong to the class of IPM (5.5). The **total variation** distance can be defined as

$$\|P - Q\|_{TV} \equiv \sup_{\|f\|_\infty \leq 1} \left| \int f \mathrm{d}P - \int f \mathrm{d}Q \right|. \tag{A.2}$$

Another important example when $(\mathcal{M}, d)$ is a Polish space is the p-Wasserstein (probability) metric (2.10). On the Wasserstein space, by theorem 6.8 [351], convergence with respect to W_p is equivalent to (and hence metrises) the convergence defined by (A.1) with respect to the space $\mathcal{C}$ of continuous functions satisfying $|f(x)| \leq C(1 + d(x_0, x)^p)$, for some $C \in \mathbb{R}, x_0 \in \mathcal{M}$.

A.3 Intuitive Description of Geometric Objects

In the following sections we will introduce elementary geometric notions in an 'intuitive" manner. These sections should thus be skipped by members of the Bourbaki school of mathematics.

A.3.1 Manifolds, Vector Fields, Flows and Differential Forms

Manifolds generalise the notions of smooth curves and surfaces to higher dimensions and are at the core of modern mathematics and physics. Simple examples include vector spaces, spheres, cylinders, and statistical manifolds. Manifolds arise by noticing that smooth geometrical shapes and physical systems are coordinate-independent concepts; so their definitions should not rely on a particular coordinate system. Coordinate patches (defined below) assign coordinates to subsets of the manifold and allow us to turn geometric questions into algebraic ones. In particular, coordinate patches allow us to transfer the calculus of $\mathbb{R}^d$ to the manifold. It is rarely possible to define a single coordinate patch over the entire manifold, except for the simplest manifolds.

A d-**dimensional manifold** is a set $\mathcal{M}$ such that every point $q \in \mathcal{M}$ has a neighbourhood $U \subseteq \mathcal{M}$ that can be described by d-coordinate functions $(x^1, \dots, x^d)$. This means that there exists a bijection $\boldsymbol{x}_U : U \to \boldsymbol{x}_U(U) \subseteq \mathbb{R}^d$, called a **coordinate patch**, which assigns the coordinates $\boldsymbol{x}_U(q) = \big(x^1(q), \dots, x^d(q)\big)$ to q. The functions $x^j : U \to \mathbb{R}$ are called **local coordinates**, which we view as being imprinted on the manifold itself. Whenever two patches $\boldsymbol{x}_U, \boldsymbol{x}_W$ overlap, $U \bigcap W \neq \emptyset$, any point q in the overlap is assigned two coordinates $\boldsymbol{x}_W(q), \boldsymbol{x}_U(q)$; in this case we require the patches to be **compatible**, i.e., the map $\boldsymbol{x}_U \circ \boldsymbol{x}_W^{-1}$, which is just the map that relates the coordinates, should be smooth (C^∞).

For example, two possible patches for the (1-dimensional) semi circle $x^2 + y^2 = 1$, $y > 0$ in a neighbourhood of $(0, 1)$ are $\boldsymbol{x}_U\big((x, y)\big) = x$, and $\boldsymbol{\theta}_U\big((x, y)\big) = \theta$ where θ satisfies $(\cos\theta, \sin\theta) = (x, y)$. The smoothness of $\boldsymbol{x}_U \circ \boldsymbol{\theta}_U^{-1} = x(\theta) = \cos\theta$ and $\boldsymbol{\theta}_U \circ \boldsymbol{x}_U^{-1} = \theta(x) = \cos^{-1} x$ implies the patches are compatible. The sphere S^2 is a 2-(sub)manifold in $\mathbb{R}^3$. In a neighbourhood of the North pole, points are specified by their x, y coordinates, since we can write z as the graph $z = z(x, y) = \sqrt{1 - x^2 - y^2}$. These points may be written as $\big(x, y, z(x, y)\big)$ and we can define a patch $\boldsymbol{x}\big(x, y, z(x, y)\big) = (x, y)$. We could have also used the spherical coordinates (θ, φ) on the upper half of S^2.

A more interesting example is the statistical manifold of Gaussian distributions $\mathcal{M} = \{\mathcal{N}(\cdot|\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_+\}$ which is a manifold endowed with global coordinates $\boldsymbol{x}_{\mathcal{M}}\big(\mathcal{N}(\cdot|\mu, \sigma^2)\big) = (\mu, \sigma^2)$ (see (5.1) for the generalisation to higher dimensional Gaussian models).

A function $f : \mathcal{M} \to \mathbb{R}$ on the manifold is said to be **smooth** at a point $q \in \mathcal{M}$ if there exists a coordinate patch $\boldsymbol{x}_U$ around q such that $f_U := f \circ \boldsymbol{x}_U^{-1} : \boldsymbol{x}_U(U) \to \mathbb{R}$ is smooth. The map f_U is just the coordinate expression of f. Since the coordinate patches are compatible, this definition of smoothness is independent of the choice of patches. The space of smooth functions on $\mathcal{M}$ is denoted $C^\infty(\mathcal{M})$.

To define Hamiltonian dynamics, we now introduce the concept of velocity of the flow of a particle on $\mathcal{M}$ (i.e., our sample space) defined by tangent vectors to the manifold. Recall that in $\mathbb{R}^d$, any vector $v = (v^1, \ldots, v^d)$ defines a directional derivative that acts on functions $f \in C^\infty(\mathbb{R}^d)$, by $v(f) \equiv \nabla_v f \equiv v \cdot \nabla f = \sum_{j=1}^d v^j \partial_j f$, where $\partial_i := \frac{\partial}{\partial x^i}$. We can thus think of the vector v as a first order differential operator $v = \sum_{j=1}^d v^j \partial_j : C^\infty(\mathbb{R}^d) \to \mathbb{R}$ (which is linear and satisfies Leibniz' rule). We now generalise this to manifolds: if $f, h \in C^\infty(\mathcal{M})$, we define a **tangent vector** $v_q : C^\infty(\mathcal{M}) \to \mathbb{R}$ at $q \in \mathcal{M}$ to be a linear map satisfying Leibniz' rule. [1]

Defining a linear combination of tangent vectors by $(au_q + bv_q)f := au_q(f) + bv_q(f)$, turns the set of tangent vectors at $q \in \mathcal{M}$ into a vector space denoted $T_q\mathcal{M}$, called the **tangent space** at q. Consider a local coordinate patch U around q. The coordinate functions x^j define tangent vectors $\partial_j|_q$ at q by [2]

$$\partial_j|_q(f) \;:=\; \left.\frac{\partial f_U}{\partial x^j}\right|_{\boldsymbol{x}(q)}.$$

These tangent vectors form a basis of $T_q\mathcal{M}$; any tangent vector at a point q is of the form $\sum_{j=1}^d v^j \partial_j|_q$, where $v^j \in \mathbb{R}$. A **vector field** v is a smooth map that assigns at each point q a tangent vector v_q. Locally any vector field can be written as $v = \sum_{j=1}^d v^j(\boldsymbol{x})\partial_j|_{\boldsymbol{x}}$ where ∂_j is the (local) vector field $\partial_j : q \mapsto \partial_j|_q$.

Vector fields are often generated by flows in mechanics. Consider particles moving on some manifold $\mathcal{M}$ (for example a fluid particle flowing along a river) which we call the **configuration manifold** (or configuration space). A mechanical system of one particle could, for example, be a mass attached at the end of a plane pendulum (so $\mathcal{M} = S^1$). The deterministic motion followed by the particle is governed by the laws of physics. Let $\Phi_t(q)$ be the position at time t of the particle that was initially at $q \in \mathcal{M}$, so $\Phi_0(q) = q$, and the trajectory followed by the particle is given by the curve $\gamma : t \mapsto \Phi_t(q)$. The curve γ generates a vector field $\dot\gamma$ over the range of γ representing the velocity of the particle; the tangent vector at the point $\gamma(a) = r$ is defined by its action on functions f:

$$\dot\gamma_r(f) \equiv \left.\frac{df(\gamma(t))}{dt}\right|_{t=a} = (f \circ \gamma)'(a).$$

Since the laws of physics are the same at all times, $\Phi_t \circ \Phi_s(q) = \Phi_{t+s}(q)$. We call Φ the **flow** and $\dot\gamma$ the **velocity field**.

The particle has a kinetic energy that measures the energy carried by its speed and mass. If no forces are acting, the particle's kinetic energy (and speed) will be constant; otherwise the force F will increase/decrease the particle's kinetic energy. Since energy is conserved, the

[1] **Leibniz' rule:** given a vector v_q at some point $q \in \mathcal{M}$, Leibniz rules is given by: $v_q(fh) = f(q)v_q(h) + h(q)v_q(f)$

[2] Technically this should be $\partial_{x^k}(f) \equiv \left.\frac{\partial f_U}{\partial r^k}\right|_{\boldsymbol{x}(p)}$ where $(r^1, \ldots, r^d)$ are (global) coordinates on $\mathbb{R}^d$

particle must be losing/gaining some other type of energy introduced by the force field, which we call potential energy U. The force is in fact an example of differential form $F = -\mathrm{d}U$, associated to the variations in potential energy. The objects $\mathrm{d}f$ or $\mathrm{d}x$ are often introduced as being mysterious "infinitesimal vectors/quantities" that give a real number when integrated. These objects are in fact special cases of **differential forms** which we shall now formally introduce: they play a central role in Hamiltonian mechanics.

A **1-form** at a point $q \in \mathcal{M}$ is a linear functional on the tangent space, i.e., a linear map $\alpha_q : T_q\mathcal{M} \to \mathbb{R}$. The simplest example is the differential of a function, $d_q f$, which maps a vector v_q to the rate of change of f in direction v_q: $d_q f(v_q) := v_q(f)$. In a coordinate patch, we can consider the differential of the coordinate function x^i. Taking $v_q = \partial_j|_q$, we see that $d_q x^i(\partial_j|_q) = \partial_j|_q(x^i) = \delta^i_j$ where δ^i_j is 1 if $i = j$ and 0 otherwise.

[3]

Hence $(d_q x^j)$ is the dual basis to $(\partial_j|_q)$ and a basis of $T^*_q\mathcal{M}$, the vector space of 1-forms at q. A **differential 1-form** α is a smooth map that assigns at each point q a 1-form α_q. Locally (i.e., in a given coordinate patch) any differential 1-form may be written as $\alpha = \sum_{j=1}^d \alpha_j(\boldsymbol{x})dx^j$ where dx^j is the (local) differential 1-form $dx^j : q \mapsto d_q x^j$. For example the differential of the function f is $df = \frac{\partial f_V}{\partial x^1}dx^1 + \cdots + \frac{\partial f_V}{\partial x^d}dx^d$.

As discussed above, a physical example of a 1-form is the force F acting on a particle, which is given by the differential of a potential energy function $F = -\mathrm{d}U$. In HMC, the potential energy U is defined as $U \equiv -\log p$ where p is the target density. Given a vector, the force measures the rate at which potential energy is gained by moving in that direction. Directions of increasing U correspond to directions of decreasing probability.

In order to defined the Riemannian metric below, we will need the tensor product of two 1-forms α, β, defined as $\alpha \otimes \beta : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$, $\alpha \otimes \beta(u_p, v_p) := \alpha(u_p)\beta(v_p)$.

A.3.2 Riemannian Metrics and Connections

In order to define what is meant by the "length" of a curve on $\mathcal{M}$, the "angle" between tangent vectors and the "volume" of a subset, it is necessary to define a metric tensor g on $\mathcal{M}$. Consider first the manifold $\mathbb{R}^3$. The standard dot product $\cdot : (v, u) \mapsto \sum_{i=1}^d v^i u^i$, applied to the basis vectors $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, $e_3 = (0, 0, 1)$ defines the Euclidean metric (or line element):

$$\mathrm{d}s^2 \;=\; G_{ij}\mathrm{d}x^i\mathrm{d}x^j \;=\; (\mathrm{d}x)^2 + (\mathrm{d}y)^2 + (\mathrm{d}z)^2,$$

[3] **Example of Differential:**
Let (θ, z) be coordinates on a cylinder. Suppose $f_V(\theta, z) := z^2 - \theta$, then $df = 2zdz - d\theta$. At $q = (1, 3)$, $d_q f = 6d_q z - d_q\theta$. The rate of change of f along $v_q = -2\partial_z|_q$ is $d_q f(v_q) = -12$

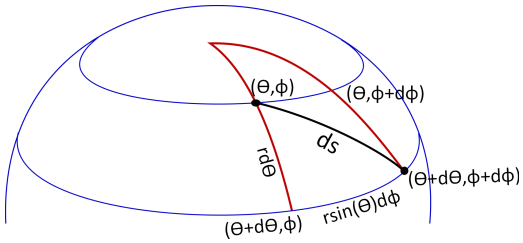

Figure A.1: Arc length on the sphere between the points (θ, ϕ) and $(\theta + d\theta, \phi + d\phi)$.

where $G_{ij} := e_i \cdot e_j$. Note that we have used the Einstein notation, where we sum over all repeated indices (i.e. $\mathrm{d}s^2 = \sum_{i,i=1}^{d} G_{ij}\mathrm{d}x^i\mathrm{d}x^j$). The line element represents the infinitesimal squared distance between $p = (x, y, z)$ and $q = (x + \mathrm{d}x, y + \mathrm{d}y, z + \mathrm{d}z)$, i.e., the squared of the length of the vector $(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z)$: $\mathrm{d}x^2 + \mathrm{d}y^2 + \mathrm{d}z^2$ (formally $(\mathrm{d}x, \mathrm{d}y, \mathrm{d}z)$ is a vector-valued 1-form as we will see below).

As an example, consider now the sphere S^2 equipped with spherical local coordinates (θ, ϕ). Since S^2 is embedded in $\mathbb{R}^3$, we have a natural definition of length between two points on the sphere: the length (as measured in $\mathbb{R}^3$) of the shortest curve on the sphere between these two points. The squared distance between two points $p = (\theta, \phi)$, $q = (\theta + \mathrm{d}\theta, \phi + \mathrm{d}\phi)$ is easily seen to be (see Fig. A.1)

$$\mathrm{d}s^2 \;=\; r^2\mathrm{d}\theta^2 + r^2\sin^2\theta\mathrm{d}\phi^2.$$

This is no longer Euclidean as the straight line between p and q does not belong in S^2. The components of the metric represent the difference between travelling on the manifold and travelling on its Euclidean representation. For example on the sphere, the components are $g_{\theta\theta} = r^2$, $g_{\phi\phi} = r^2\sin^2\theta$ and $g_{\theta\phi} = 0$, which shows that when we travel an amount $\mathrm{d}\phi$ and $\mathrm{d}\theta$ on the coordinate space, we travel a distance squared $r^2\sin^2\theta\mathrm{d}\phi^2$ and $r^2\mathrm{d}\theta^2$ respectively on the manifold. The fact that $g_{\theta\phi} = 0$ shows the coordinates lines (lines of constant ϕ or θ) are orthogonal. [4] We could have of course used a different system of local coordinates to express the metric, for example $\mathrm{d}s^2 = dl^2 + r^2\sin^2(\frac{l}{r})\mathrm{d}\phi^2$, where l is the distance to the point from the north pole along a great circle. As mentioned above, the metric characterises metric quantities (such as lengths, volumes, angles, etc...) which are intrinsic properties of the manifold. Hence the metric must depend only on the manifold and not on the choice of local coordinates on it. For example, it is impossible to find coordinates on the sphere such that ds^2 is Euclidean, because the sphere is curved.

The notation $\mathrm{d}s^2$ arises from the fact that classically, we think of $\mathrm{d}s$ as the "infinitesimal

[4] For those familiar with continuum mechanics, we may think of ds^2 as a deformation: let $\boldsymbol{r} : U \subset \mathbb{R}^2 \to S^2 \subset \mathbb{R}^3$ be the embedding of the sphere: $\boldsymbol{r}(\theta, \phi) = (\sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta)$. Then ds^2 is the "deformation tensor" (also called right Cauchy-Green tensor in continuum mechanics), which defines how lines on the subset U are deformed (up to rigid motion) when mapped onto the sphere .

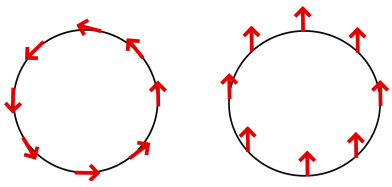

Figure A.2: LHS: constant vector field wrt $\mathrm{d}r^2 + r^2\mathrm{d}\theta^2$. RHS: constant vector field wrt $\mathrm{d}r^2 + \mathrm{d}\theta^2$

line element" of the manifold embedded in $\mathbb{R}^d$. Mathematically $\mathrm{d}s$ is actually a linear transformation, which maps a tangent vector on $\mathcal{M}$ to itself, and views it as a vector in $\mathbb{R}^d$, i.e., $\mathrm{d}s : T_p\mathcal{M} \to T_p\mathbb{R}^d$ with $\mathrm{d}s(v) = v$.[5] The line element is then defined to be the quadratic form $\mathrm{d}s^2(v) := \mathrm{d}s(v) \cdot \mathrm{d}s(v)$, which defines the length of a vector v to be its length in $\mathbb{R}^d$.

Let us now define $\mathrm{d}s^2$ formally (without referring to the embedding space). The line element $\mathrm{d}s^2$ is in fact a (Riemannian) metric, i.e. a smooth map g which assigns to each $p \in \mathcal{M}$ an inner product $g_p : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$. In local coordinates (x^j), recall the differential $\mathrm{d}_p x^j : T_pU \to \mathbb{R}$, maps a vector $v_p = v^k \partial_{x^k}|_p$ to $\mathrm{d}x^j(v_p) = v^j$. Defining $G_{ij} := g(\partial_{x^i}, \partial_{x^j})$, it is easy to check the metric may be locally written in terms of these differentials:

$$g \;=\; G_{ij}\mathrm{d}x^i \otimes \mathrm{d}x^j.$$

Any manifold equipped with a Riemmanian metric is called a Riemmanian manifold. Let $\gamma : (a, b) \to \mathcal{M}$ be a curve on the manifold $\mathcal{M}$. Recall that its length is simply given by integrating the speed over time (since velocity = space/time). The velocity of γ is its tangent vector $\dot{\gamma}$ and the speed is the length of velocity $\sqrt{g(\dot{\gamma}, \dot{\gamma})}$, the length of γ is thus $\int_a^b \sqrt{g(\dot{\gamma}(t), \dot{\gamma}(t))}dt$. This enables us to turn $\mathcal{M}$ into a metric space, by defining the (Riemann) distance between two points p, q to be the shortest curve between them

$$d(p, q) \;:=\; \inf_{\gamma:\gamma(0)=p,\gamma(1)=q} \int_0^1 \sqrt{g(\dot{\gamma}(t), \dot{\gamma}(t))}dt.$$

Let $\Gamma(T\mathcal{M})$ be the space of vector fields on $\mathcal{M}$. If you travel with constant velocity on a sphere (say with velocity field ∂_ϕ), you are not travelling with constant velocity in $\mathbb{R}^3$, since your direction of travel keeps varying (in other words, ∂_ϕ is a constant vector field on S^2 but $\mathrm{d}s(\partial_\phi)$ is a varying vector field on $\mathbb{R}^3$); see Fig. A.2 for an illustration on the circle S^1. Defining the curvature of the manifold is equivalent to defining an intrinsic notion of rate of change of vector fields over $\mathcal{M}$ (more precisely, defining which vector fields are constant on $\mathcal{M}$ defines the curvature of $\mathcal{M}$).

Defining the rate of change of vector fields on $\mathcal{M}$ is more complicated than defining the

[5] In other words, if $i : \mathcal{M} \to \mathbb{R}^d$ is the inclusion map, then $\mathrm{d}s$ is the pushforward i_*.

rate of change of functions in $C^\infty(\mathcal{M})$, since these have a fixed codomain $\mathbb{R}$ making it easy to compare the values of the function at different point. On the other hand, vector fields assign to each point $p \in \mathcal{M}$ a vector in a distinct vector space $T_p\mathcal{M}$. Let v be a vector field in $\mathbb{R}^3$ (i.e. $v \in \Gamma(T\mathbb{R}^3)$). Then clearly v is constant on $\mathbb{R}^3$ if and only if it is constant along any curve γ;

$$D_{\dot\gamma} v \ := \dot\gamma \cdot \nabla v = \dot\gamma^i \frac{\partial v^k}{\partial x^i} \partial_{x^k} = \ \frac{\mathrm{d}v(\gamma(t))}{dt} \ = \dot v = \ 0.$$

If v is a vector field on a surface $\mathcal{M}$ in $\mathbb{R}^3$ and γ a curve in $\mathcal{M}$, the (intrinsic) rate of change is given by the rate of change of v as a vector field in $\mathbb{R}^3$ minus the rate of change of v along the normal vector field n to $\mathcal{M}$:

$$D_{\dot\gamma} v \ := \ \dot\gamma_* \cdot \nabla v_* - \Big(\big(\gamma_* \cdot \nabla v_*\big) \cdot n \Big) n, \tag{A.3}$$

where $v_* := \mathrm{d}s(v)$ and $\dot\gamma_* := \mathrm{d}s(\dot\gamma)$. More generally, the curvature is characterised by a connection i.e., a map $D : \Gamma(T\mathcal{M}) \times \Gamma(T\mathcal{M}) \to \Gamma(T\mathcal{M})$ such that:

$$\begin{aligned} D_v(u+j) &= D_v u + D_v j, \\ D_v(fu) &= v(f)u + f D_v u, \\ D_{u+v} j &= D_u j + D_v j, \\ D_{fv} u &= f D_v u, \end{aligned}$$

where $D_v u := D(v,u)$, represents the rate of change of u along v, or the covariant derivative of u along v. To understand the meaning of the connection, it is useful to look at its local expression. Let (∂_{x^k}), be a local basis of vector fields, we define the components of the vector potential[6] A^i_{lk} of the connection by

$$D_{\partial_{x^k}} \partial_{x^l} \ = \ A^i_{kl} \partial_{x^i}.$$

Then it is easy to check

$$D_v u \ = \ v^r \big(\partial_r u^i + A^i_{rl} u^l \big) \partial_i,$$

which is a straightforward generalisation of Eq (1). The first term yields the standard directional derivative of vector fields in $\mathbb{R}^d$ and the vector potential components "project" this derivative onto the manifold. We can define many connections on any manifold $\mathcal{M}$, however every Riemannian manifold has a unique connection that is symmetric and compatible with the metric, and corresponds to the usual notion of connection on surfaces in $\mathbb{R}^3$. This is called

[6] Formally the vector potential over a neighbourhood U is a $\mathrm{End}(TU)$-valued 1-form on U, or in other words a section of $\mathrm{End}(TU) \otimes T^*U$. Here $\mathrm{End}(TU)$ stands for the space of endomorphisms of TU.

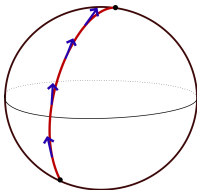

Figure A.3: The geodesics on S^2 are the great circles. The tangent vector to a great circle, when viewed as a vector in $\mathbb{R}^3$, varies only along the normal to S^2.

the Levi-Civita connection. To define it we need to extend D_v to tensor fields[7]. Define D_v on functions by $D_v f := v(f)$, and on differential forms by assuming a Leibniz rule on the contraction $(\omega, u) \mapsto \omega(u)$:

$$\big(D_v(\omega)\big)(u) \;:=\; v\big(\omega(u)\big) - \omega\big(D_v u\big),$$

and to tensor fields T, S, by assuming linearity and a Leibniz rule for $\otimes$:

$$D_v(T \otimes S) \;:=\; (D_v T) \otimes S + T \otimes D_v S.$$

The Levi-Civita connection on $(\mathcal{M}, g)$ is the unique connection which is symmetric (i.e. $A^i_{jk} = A^i_{kj}$) and preserves the inner product of vectors that are transported through a constant move (i.e., in such a way to preserve their direction on magnitude) on $\mathcal{M}$, that is if u_p, v_p are vectors at p which are constantly transported to q, then $g(u_p, v_p) = g(u_q, v_q)$ (mathematically this requires that $D_v g = 0$, $\forall v$). The components of the vector potential of the Levi-Civita connection are called Christoffels symbols. They are usually denoted by Γ^i_{jk}.

A geodesic is a path γ which is as straight as possible at every step of the way. For a surface embedded in $\mathbb{R}^3$ this means that $\dot\gamma$ is parallel to the normal field, i.e., the tangent vector field varies only in the direction normal to the surface (fig with zero geodesic curvature). More generally this means $D_{\dot\gamma}\dot\gamma = 0$, which in local coordinates gives

$$\frac{d^2\gamma^i}{dt^2} + A^i_{jk}\frac{d\gamma^j}{dt}\frac{d\gamma^k}{dt} \;=\; 0.$$

[7]Recall a tensor field is a $C^\infty(\mathcal{M})$-multilinear map $\Gamma(T\mathcal{M}) \times \cdots \times \Gamma(T\mathcal{M}) \times \Gamma(T^*\mathcal{M}) \times \cdots \times \Gamma(T^*\mathcal{M}) \to C^\infty(\mathcal{M})$, where $\Gamma(T^*\mathcal{M})$ is the space of differential forms on $\mathcal{M}$.

A.4 Further Background Material

A.4.1 Origins of HMC

Consider the Dirac Lagrangian of QCD [150]

$$\mathcal{L}_D = \Re\left(i\psi_r^\dagger \gamma^\mu \partial_\mu \psi_r + i\psi_r^\dagger \gamma^\mu A_\mu \psi)_r)\right) + m\psi_r^\dagger \psi_r$$

where the spinor field and gauge bosons are described by maps $\psi : \mathbb{R}^4 \to \mathbb{C}^4 \otimes \mathbb{C}^3$, and $A \in \Omega^1(\mathbb{R}^4, \mathfrak{su}(3))$, γ^μ are the (Hermitian) Dirac gamma matrices. We discretise the QCD gauge theory by constructing the discrete analogue of the Dirac and Yang-Mills action on a lattice [361], and in particular the lattice QCD partition function, which is the object of interest in QCD-HMC. Spacetime $\mathcal{M} = \mathbb{R}^4$ is replaced by a lattice with spacing a, and the fermionic matter fields ψ are discretised by viewing them as Grassmann fields on the lattice (which output a Grassmann variable on each lattice site). To discretise the gauge field A we proceed as follows: consider a site x on the lattice and the curve $\gamma : [0,1] \to \mathbb{R}^4$ from x to its nearest neighbour $x + a\partial_\mu$ in direction ∂_μ. There exists a unique horizontal lift (with respect to $\mathcal{A}$) $\gamma^L : [0,1] \to \mathcal{F}$ of γ s.t. $\gamma^L(0) = s(x)$ where $\mathcal{F}$ is a principal SU(3) bundle over $\mathcal{M}$ (this is the parallel transport in $\mathcal{F}$ along γ). We can find a unique curve $g : [0,1] \to \mathrm{SU}(3)$ for which $\gamma^L(t) = g(t) \cdot s\left(\gamma(t)\right)$. The group element $g(1)$ is called the **link variable** and denoted by $U_\mu(x)$ in the literature (where it is usually expressed as a path-ordered exponential) and the set of these elements (varying direction μ and site x) will represent the discretised gauge field (see 5.8 and 5.10 [150] for a discussion on parallel transport and path-ordered exponentials). The link variables satisfy $U_{-\mu}(x) = U_\mu(x - a\partial_\mu)^\dagger$ (see 5-6 [144] for a discussion on discretisation and relation to statistical mechanics).

Using the discretised fields we can discretise the actions. The Yang-Mills action (associated to the connection) is replaced by the **Wilson gauge (or plaquette) action**

$$\mathcal{S}_{WG} \equiv \beta \sum_x \sum_{\mu<\nu} \left(1 - \frac{1}{3}\Re \operatorname{tr}[P_{\mu\nu}(x)]\right)$$

where $P_{\mu\nu}(x) \equiv U_\mu(x) U_\nu(x + a\partial_\mu) U_\mu^{-1}(x + a\partial_\mu) U_\nu^{-1}(x)$ is called the **plaquette**, and $\beta \geq 0$ is a parameter related to the **gauge coupling** $g_c > 0$, which is introduced by rescaling the connection 1-form, and β is chosen to ensure that in the continuum limit $a \to 0$ the Wilson gauge action $\mathcal{S}_{WG}$ converges to the Yang-Mills action. The discretisation of the Dirac action is a much more complicated procedure, it can in fact be shown that it is not possible to discretise the

action without breaking at least one of: chiral symmetry, convergence to the correct continuum limit, locality of the Dirac operator, or introducing a doubling fermion problem. One option is the **Wilson fermion action**

$$\mathcal{S}_{WM} \equiv a^4 \sum_x \psi^\dagger(x) \left(\gamma^\mu \Delta_\mu + a \sum_\mu \Box_\mu + m \right) \psi(x) \equiv \sum_x \psi^\dagger(x) D\psi(x)$$

where $\gamma^\mu \Delta_\mu + a \sum_\mu \Box_\mu$ is the **Wilson-Dirac operator** where

$$\Delta_\mu \psi(x) \equiv \frac{U_\mu(x)\psi(x + a\partial_\mu) - U_\mu^{-1}(x - a\partial_\mu)\psi(x - a\partial_\mu)}{2a}$$

is the finite-difference covariant operator and the finite-difference Laplacian

$$\Box_\mu \psi(x) \equiv \frac{U_\mu(x)\psi(x + a\partial_\mu) - 2\psi(x) + U_\mu^{-1}(x - a\partial_\mu)\psi(x - a\partial_\mu)}{2a^2}$$

is introduced to remove the fermion doublets. The resulting action $\mathcal{S}_{WM}$ has the right properties except it violates chiral symmetry (see 2.3 [157]).

Given all these definitions, we can write down the partition function of discretised QCD which will lead us to the original Hamiltonian of HMC. The partition function is defined as

$$Z \equiv \int \prod_{x,\mu,f} \mathrm{d}U_\mu(x) \mathrm{d}\psi_f^\dagger(x) \mathrm{d}\psi_f(x) e^{-S_{WG} + \sum_f S_{WM}^f}$$

where $S_{WM}^f \equiv \sum_x \psi_f^\dagger(x) D_f \psi_f(x)$ is the Wilson fermion action corresponding to a fermion ψ_f with **flavour** f, $\mathrm{d}U_\mu(x)$ denotes the Haar measure, and $\mathrm{d}\psi_f^\dagger(x)\mathrm{d}\psi_f(x)$ represents a Grassmann integral. Using properties of Grassmann integrals the partition function becomes

$$Z = \int \prod_{x,\mu} \mathrm{d}U_\mu(x) \prod_f \det D_f e^{-S_{WG}}.$$

If we have two flavours ($f = 2$) of degenerate (same) mass, then $\prod_f \det D_f = \det D \det D = \det DD^\dagger \propto \int \mathrm{d}\phi^\dagger \mathrm{d}\phi e^{-\phi^\dagger (DD^\dagger)^{-1} \phi}$ where we have introduced the pseudofermion (boson) ϕ using a Gaussian functional integral. If we further introduce some fictitious momenta $\mathfrak{su}(3)$-valued $p_{x,\mu}$ variables on the links, the partition function is finally

$$Z = \int \prod_{x,\mu} \mathrm{d}U_\mu(x) \mathrm{d}p_{x,\mu} \mathrm{d}\phi^\dagger \mathrm{d}\phi e^{-S_{WG} - \phi^\dagger (DD^\dagger)^{-1} \phi - \frac{1}{2} \sum_{x,\mu} \langle p_{x,\mu}, p_{x,\mu} \rangle_{\mathfrak{su}(3)}}.$$

Hence, we have obtained the Hamiltonian of interest in QCD simulations (see [332, 14] for a

discussion on the implementation of HMC with an odd number of flavors)

$$H = S_{WG} + \phi^\dagger (DD^\dagger)^{-1}\phi + \tfrac{1}{2}\sum_{x,\mu} \langle p_{x,\mu}, p_{x,\mu}\rangle_{\mathfrak{su}(3)} \,,$$

and a typical HMC-QCD algorithm proceeds as follows

- refresh the momentum $p_{x,\mu} \sim \mathcal{N}(0, I)$ (assuming the basis of $\mathfrak{su}(3)$ is orthonormal) on every link (x, μ) of the lattice in parallel
- pseudofermion refreshment $\phi \sim \mathcal{N}(0, DD^\dagger)$. This may be done by sampling $\eta \sim \mathcal{N}(0, I)$ and setting $\phi \equiv D\eta$.
- Update the link variable and momenta on every link in parallel using an appropriate integrator and a Monte Carlo correction step with effective potential $V_{eff} \equiv S_{WG} + \phi^\dagger (DD^\dagger)^{-1}\phi$. At the current state (p, ϕ, U), the kinetic flow preserves (p, ϕ) and maps $U \mapsto Ue^{tp}$, while the potential flow preserves (ϕ, U) and maps $p \mapsto p + \delta t \mathrm{d}V_{eff}$.

During such simulations most of the computational work arise in the calculation of the derivative of the pseudofermion term:

$$\frac{\partial \phi^\dagger (DD^\dagger)^{-1}\phi}{\partial U} = -\phi^\dagger (DD^\dagger)^{-1} \frac{\partial DD^\dagger}{\partial U} (DD^\dagger)^{-1}\phi$$

since this requires solving the linear system $DD^\dagger v = \phi$ (see II.E [179]). It is thus common to separate the potential energy into $\phi^\dagger (DD^\dagger)^{-1}\phi$ and S_{WG} and use shorter time steps for the gauge action and larger time steps for the pseudofermion term during the numerical integration, as discussed above [315]. This results in an improved acceptance rate compared to leapfrog due to the more accurate integration of the (less expensive) gauge action. This can be combined with other methods to speed-up HMC, such as [152] which proposes an alternative integral expansion of $\det DD^\dagger$ by splitting it into two factors $\det DD^\dagger = \det\left(\tilde{D}\tilde{D}^\dagger\right) \det\left(\tilde{D}^{-\dagger} DD^\dagger \tilde{D}^{-1}\right)$, and use an integral pseudofermions expansion for each determinant. Choosing $\tilde{D}$ with a smaller condition number than D allows for larger step sizes to be used at constant acceptance rate.

The two-stage integrator $\Phi^V_{\lambda\delta t} \circ \Phi^T_{\frac{1}{2}\delta t} \circ \Phi^V_{(1-2\lambda)\delta t} \circ \Phi^T_{\frac{1}{2}\delta t} \circ \Phi^V_{\lambda\delta t}$ also plays an important role in lattice QCD, where it was found to be 50% more efficient than the leapfrog integrator [333]. The integrator has a free parameter λ which was tuned in [284] to the value λ_c that minimises the norm of the coefficients of the Poisson brackets in the second-order term in (3.14), giving $\lambda_c \equiv \min_\lambda \sqrt{\left(\frac{6\lambda^2-6\lambda+1}{12}\right)^2 + \left(\frac{1-6\lambda}{24}\right)^2} \approx 0.1931833275$. When λ is set to λ_c, this is also referred to as the **second order minimum norm integrator**. In [333] a more accurate tuning was proposed by giving different weights (obtained from numerical simulations) to the coefficients of the Poisson brackets resulting in a slightly larger value for λ_c. The optimal parameter λ_c

may viewed as minimising[8]

$$\langle (\mathcal{H}_\Psi - H)^2 \rangle \equiv \mathbb{E}[(\mathcal{H}_\Psi - H)^2] \equiv \frac{1}{Z} \int (\mathcal{H}_\Psi - H)^2 e^{-H} \omega^n,$$

where $\mathcal{H}_\Psi$ is the shadow Hamiltonian (see §3.5.4).

An alternative **force-gradient integrator** based on the fact that the flow of the Hamiltonian vector field $\{V, \widehat{\{V, T\}}\}$ can be computed for the standard Hamiltonians of lattice QCD was proposed in [189] (here $e^{\delta t \hat{H}} \equiv \Phi_{\delta t}^{X_H}$)

$$\Psi_{\delta t} \equiv e^{\frac{1}{6}\delta t \hat{V}} e^{\frac{1}{2}\delta t \hat{T}} e^{\frac{48\delta t \hat{V} - \delta t^3 \widehat{\{V,\{V,T\}\}}}{72}} e^{\frac{1}{2}\delta t \hat{T}} e^{\frac{1}{6}\delta t \hat{V}}.$$

The name arises from the fact that $\{V, \widehat{\{V, T\}}\}$ involves second derivatives of the potential energy, which correspond to a force gradient (since the force is a first derivative of the potential energy). The additional force-term $\{V, \widehat{\{V, T\}}\}$ increases the computational cost of the integrator, but it is expected to be more efficient than second-order minimum norm integrators for large lattices [86]. It was shown in [189] that the force-gradient integrator outperform Campostrini integrators, and that its shadow Hamiltonians approximates the true Hamiltonian to fourth order. Indeed its parameters were specifically chosen to eliminate the coefficients of $O(\delta t^2)$ terms (although this is not necessarily optimal for HMC, where one could also instead adapt the parameters to the model at hand to minimise the cost of integrator).

A.4.2 Shadow HMC

One of the benefits of HMC compared to many MCMC methods is that distant proposals can still be accepted at a satisfying rate since the energy does not change drastically along the numerical trajectories of symplectic integrators, which allows HMC to generates less correlated samples. However the acceptance rate decreases exponentially as the dimension of the sample/configuration space increases (the integrator has order ν) [302]:

$$\mathbb{E}[\delta H] = \int \delta H e^{-H} \omega^n = O(n \delta t^{2\nu}).$$

It is thus reasonable to wonder if we can make use of the Shadow Hamiltonian, which is exactly preserved by the symplectic integrator, to improve the acceptance rate. This is the aim of the **shadow HMC** (SHMC) methods, which were originally proposed in [177]. An immediate difficulty arises from the fact that the shadow Hamiltonian is (usually) only known as an

[8] Under the assumption that δt is small, so we can restrict to the terms of order δt^2, and the Poisson brackets are equal.

asymptotic expansion (see §3.5.4.1), so we need to work with a truncation. If $\mathcal{H}^k_{\delta t}$ denotes the truncation up to order δt^k of the shadow Hamiltonian (where $k > \nu$), we have (2.2.1 [302])

$$\mathbb{E}[\delta \mathcal{H}^k_{\delta t}] = O(n\delta t^{2k})$$

which leads to an improved acceptance rate if $\mathcal{H}^k_{\delta t}$ is used in the accept/reject step. Methods based on the shadow HMC proceed as follows. Given a target measure $P \propto e^{-H}\omega^n$, and a symplectic time-reversible symmetric integrator with fixed step-size δt,[9] we compute its associated truncated shadow Hamiltonian and define a modified Hamiltonian $\tilde{\mathcal{H}}$ based on it. We then sample from the modified target measure $\tilde{P} \propto e^{-\tilde{\mathcal{H}}}\omega^n$ by

- partially or fully refreshing the momentum in a way that preserves $\tilde{P}$, which usually requires and accept/reject step,
- compute the numerical trajectory using the integrator, and accept proposed state with probability $\min\left(1, e^{-\delta\tilde{\mathcal{H}}}\right)$; if rejected flip the momentum.
- Once samples (q^i, p_i) are obtained, in order to correct the bias introduced from the fact these are samples from $\tilde{P}$ rather than P we compute the weights

$$w_i \equiv e^{-\left(H(q^i,p_i)-\tilde{\mathcal{H}}(q^i,p_i)\right)}$$

and estimate $\int f(q,p)\mathrm{d}P$ using the self-normalised importance sampling estimator[10]

$$\frac{\sum_i f(q^i,p_i)w_i}{\sum_i w_i}.$$

The original SHMC method [177] relied on the standard leapfrog integrator for a Hamiltonian $H(q,p) = \frac{1}{2}p^\top G^{-1}p + V(q)$. To compute the terms in the shadow Hamiltonian expansion it used a procedure proposed in [319] which works for symplectic integrators obtained from the splitting method. The algorithm however introduces a new parameter controlling the acceptance rate of the momentum refreshment and is computationally expensive. To remedy these shortcomings [331] introduced **separable shadow HMC** (S2HMC), which relies instead on a processed leapfrog integrator constructed using the method of generating functions. The processed leapfrog integrator has the advantage that the second-order expansion of its shadow Hamiltonian is now separable, with kinetic energy $\frac{1}{2}p^\top G^{-1}p$ (and a modified potential energy). The momentum refreshment step then simply samples from a Gaussian and no extra parameter is introduced.

[9] See IX.6 [148] for a discussion on variable time-steps

[10] Recall this estimator follows from $\int f(q,p)\mathrm{d}P = \int f(q,p)\frac{e^{-H}/Z_P}{e^{-\tilde{H}}/Z_{\tilde{P}}}\mathrm{d}\tilde{P} = \frac{\int f(q,p)\frac{Z_{\tilde{P}}e^{-H}}{Z_P e^{-\tilde{H}}}\mathrm{d}\tilde{P}}{\int \frac{Z_{\tilde{P}}e^{-H}}{Z_P e^{-\tilde{H}}}\mathrm{d}\tilde{P}} = \frac{\int f(q,p)\frac{e^{-H}}{e^{-\tilde{H}}}\mathrm{d}\tilde{P}}{\int \frac{e^{-H}}{e^{-\tilde{H}}}\mathrm{d}\tilde{P}}$.

In §3.4.2.2 we will discuss HMC with partial momentum refreshment (usually called GHMC) and how it can improve HMC by requiring fewer leapfrog steps, provided the acceptance rate is sufficiently high. The **generalised shadow HMC** (GSHMC) [8] aims to achieve this latter requirement by combining partial momentum refreshments with shadow Hamiltonians. Moreover GSHMC allows the use of generalised coordinates and Riemannian (non-constant) kinetic energies - the Hamiltonian can take the form $H(q,p) = \frac{1}{2}p^\top G^{-1}(q)p + V(q)$. The symplectic integrator considered in GSHMC is the generalised leapfrog method (§A.5.5.1) constructed using the discrete variational principle. Truncated shadow Hamiltonians are obtained by substituting a Taylor expansion for the discrete curve in the discrete Lagrangian which is then truncated, viewed as a Lagrangian function of higher-order derivative, and transformed into the corresponding Hamiltonian (see II [318]). GSHMC then proceeds as in GHMC with the truncated shadow Hamiltonian (which requires an additional accept-reject step after the partial momentum refreshment, and a reweighing procedure once the samples have been obtained).

A.4.3 Approximate-Gradient HMC for Statistical Applications

One of the drawbacks of the standard HMC algorithm 2 is that it requires evaluating the potential energy (during the accept/reject step) and its derivative (during the leapfrog integration), which are often computationally expensive in high dimensional problems. For example in the Bayesian paradigm the aim is to sample from the posterior distribution $\rho(q|D)\mathrm{d}q$ associated to a set of i.i.d. observations D. In this common scenario the potential energy is given by

$$V(q) \equiv -\sum_{x_i \in D} \log \rho(x_i|q) - \log \rho(q), \tag{A.4}$$

where $\rho(q)$ is the prior density, $\rho(x_i|\cdot)$ the likelihood. Evaluating V and its gradient is then typically challenging due to the large numbers of observations contained in D.

Several methods have been proposed to alleviate the computational burden by resorting to approximations of the potential energy, or of its gradient, during the implementation of HMC. Note that HMC does not require the MD step to be based on the "target" Hamiltonian H: any surrogate Hamiltonian can be used. However if a surrogate Hamiltonian is used during both the MD and MC steps, then the algorithm does not preserve the target distribution, i.e., it generates biased samples. On the other hand, if the MD step uses a surrogate Hamiltonian, while the MC step uses the true Hamiltonian, then the algorithm samples without bias. But then the conservation of acceptance rate (see §3.5.4.3), resulting from the fact the discretised mechanical system preserves the shadow Hamiltonian which stays close to the true Hamiltonian for exponentially long time, will typically no longer hold.

Let us now discuss some examples where approximate Hamiltonians are used. The method in [369] is based on finding a bounded region $R \subset \mathcal{M} = \mathbb{R}^n$ (in practice a hypercube) containing most of the probability mass of the target distribution. A surrogate gradient is then formed: over R the gradient is approximated using Smolyak algorithm (a sparse grid interpolation method), while the true gradient is used elsewhere. The leapfrog integration is then implemented using the surrogate gradient. In [368] a surrogate is defined by $V \approx \sum_{i=1}^{s} v_i \sigma(\cdot; \gamma_i)$, where σ is some specified "activation" function, and s is the number of "neurons". The approximation is constructed using extreme learning machine: the parameters (γ_i) are sampled from some auxiliary distribution, and are then used to train a random network to obtain the parameters (v_i). The training data for the network is given by samples from the target distribution obtained during the burn-in phase of a standard HMC, together with evaluations of V at these samples. Once the surrogate for V is defined, it is used during the leapfrog integration. In [370] instead the parameters (v_i) are constructed by minimising (a regularised empirical version of) the score-matching criterion $(v_i) \mapsto \mathrm{SM}(Q[(v_i)] \| P)$ where $Q([v_i]) \propto \exp\left(-\sum_{i=1}^{s} v_i \sigma(q; \gamma_i)\right) \mathrm{d}q$. The proposed sampling procedure uses a HMC algorithm with potential energy given by a regularisation (based on the Hessian of the true potential energy) of the surrogate function, and the parameters (v_i) of the surrogate are updated as new samples are generated (note the samples obtained will be biased).

The **Stochastic gradient HMC** (SGHMC) algorithm [77] aims to reduce the cost of HMC by replacing the sum over all data points in (A.4) by a sum over a uniformly sampled minibatch. One then obtains an estimator $\nabla_q \tilde{V}$ which is assumed to satisfy

$$\nabla_q \tilde{V} \approx \nabla_q V + \sigma(q) W$$

where $W \sim \mathcal{N}(0, I)$, and $\sigma(q)$ is a matrix associated to the noise. The idea is to then replace $\nabla_q V$ in Hamilton's equations with $\nabla_q \tilde{V}$. The resulting stochastic dynamics does not preserve the canonical distribution, and it is necessary to add a "friction term" which yields the Langevin dynamics of (3.9). However the noise covariance $\sigma\sigma^\top$ is unknown and must be estimated, which led the authors to further modify the equations of motion and obtain a dynamics that does not preserve the target distribution. The SGHMC algorithm is constructed by discretising this dynamics (with no Metroplis correction), based on the argument that for time-steps converging to 0, the gain in efficiency outweighs the loss in accuracy. It is argued in [42] that it is in fact not possible to maintain the scalable and robust properties of HMC when using minibateches, as approximate gradients generate trajectories that no longer maintain the good energy-preserving properties of symplectic integrators, leading to highly biased expectations.

In [330] it was proposed to use exponential families to approximate target density and its derivative in cases where these were intractable. Given a (unnormalised) target density ρ :

$\mathbb{R}^n \to \mathbb{R}$, the idea is to find a parameter θ in an exponential family modelled over a reproducing kernel Hilbert space $\mathcal{H}_k$ (RKHS); that is we want to find a $\theta^* \in \Theta \subset \mathcal{H}_k$ s.t., $\rho \approx p_{\theta^*}$ and then approximate the target log-gradient by $\nabla_x \log \rho \approx \nabla_x \log p_{\theta^*} = \nabla_x \theta^*$. The approximation is found by first noticing we can rewrite the score-matching divergence as [11]

$$\mathrm{SM}(P\|P_\theta) = \int \left(\Delta_x \log p_\theta + \tfrac{1}{2}\|\nabla_x \log p_\theta\|_2^2\right) \mathrm{d}P(x) + \text{constant}$$

where the constant does not depend on θ. Given a i.i.d sample $\{\omega_j\}_{j=1}^m$ from ρ, an optimisation problem is then solved to find the value θ^* which minimises the empirical estimate $\frac{1}{m}\sum_{j=1}^m \left(\Delta_{w_j} \log p_\theta + \frac{1}{2}\|\nabla_{w_j} \log p_\theta\|_2^2\right)$ of the function $\theta \mapsto \mathrm{SM}(P\|P_\theta)$. Once θ^* is given, the algorithm proceeds as a standard HMC scheme, except that proposals are generated using a volume-preserving reversible integrator which approximates the flow of a Hamiltonian H^* formed by replacing $\log \rho$ with θ^*.

A.4.4 Hilbert Space HMC

Typically the performance of HMC on the Euclidean space $\mathbb{R}^n$ deteriorates as n increases, in the sense that the number of MCMC steps necessary to approximate the target distribution grows with n. To overcome this curse of dimensionality, it was proposed in [40] to construct an HMC algorithm that is well-defined on an infinite dimensional space $\mathscr{H}$, and view the target distribution on the very high-dimensional space $\mathbb{R}^n$ as a finite-dimensional of a measure on $\mathscr{H}$.

More precisely consider a target distribution P on a separable Hilbert space $\mathscr{H}$ assumed to be absolutely continuous with respect to a Gaussian reference measure $\mathcal{N}(0, \mathcal{C})$, where $\mathcal{C}$ is the (bounded linear) covariance operator on $\mathscr{H}$, $\mathcal{C}(\cdot) \equiv \mathbb{E}[Id \otimes Id] = \int_{\mathscr{H}} x \langle x, \cdot \rangle_{\mathscr{H}} \mathcal{N}(0, \mathcal{C})(\mathrm{d}x)$. Distributions on Hilbert spaces occur in variety of applications, such as Bayesian inverse problems. The potential energy, denoted $\Phi : \mathscr{H} \to \mathbb{R}$, is by definition

$$\frac{\mathrm{d}P}{\mathrm{d}\mathcal{N}(0, \mathcal{C})} \propto e^{-\Phi},$$

and the measure is then lifted to the "canonical distribution" $P_{can} \propto e^{-\Phi}\mathcal{N}(0, \mathcal{C}) \otimes \mathcal{N}(0, \mathcal{C})$ on $\mathscr{H} \times \mathscr{H}$. In the finite dimensional case this corresponds to a target $P(\mathrm{d}q) \equiv e^{-\Phi(q)} e^{-\frac{1}{2} q^\top \mathcal{C}^{-1} q} \mathrm{d}q$, a kinetic energy $T(v) \equiv \frac{1}{2} v^\top \mathcal{C}^{-1} v$, and a canonical distribution

$$P_{can}(\mathrm{d}q, \mathrm{d}v) \equiv \frac{1}{Z} e^{-H(q,v)} \det(\mathcal{C}^{-1}) \mathrm{d}q \mathrm{d}v \propto e^{-\Phi(q)} e^{-\frac{1}{2} q^\top \mathcal{C}^{-1} q} e^{-\frac{1}{2} v^\top \mathcal{C}^{-1} v} \mathrm{d}q \mathrm{d}v,$$

whose Hamilton's equations are $\dot{q} = v$, $\dot{v} = -q - \mathcal{C}\nabla\Phi$ from (3.6). These equations still make

[11] Provided Stoke's theorem holds without boundary terms and the densities are sufficiently differentiable

sense in infinite dimension, and it is shown that their flow preserve P_{can} (see theorem 3.2 [40]). An HMC scheme is then constructed using a numerical integrator based on the split into the flow of the force term

$$\dot{q} = 0, \quad \dot{v} = -\mathcal{C}\nabla\Phi, \quad \Rightarrow \quad q(t) = q(0), \quad v(t) = v(0) - t\mathcal{C}\nabla\Phi,$$

and the harmonic term

$$\dot{q} = v, \quad \dot{v} = -q, \quad \Rightarrow \quad q(t) = \cos(t)q + \sin(t)v, \quad v(t) = \cos(t)v - \sin(t)q$$

together with a Metropolis step.

A.4.5 Ergodicity of HMC

Recall that the aim of HMC is to approximate $\mathbb{E}[f] = \int f(x)\mathrm{d}P(x)$, for some integrable observable f, with the random quantity $\frac{1}{m}\sum_{i=1}^{m} f(X_i)$, where $X_i : \Omega \to \mathbb{R}^n$ are the random variables of the Markov chain generated by the HMC algorithm targeting P. We say the chain is **pathwise ergodic** if for P-almost all initial values $X_1 = x$, we have

$$\lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} f(X_i) \overset{\text{a.s.}}{=} \int f(x)\mathrm{d}P(x),$$

which holds for example if (X_i) is aperiodic and P-irreducible (proposition 2.1 [329]). While HMC is commonly ergodic this is not guaranteed in general. To address this issue it is usually sufficient randomise the trajectory length of the dynamics [244, 43]. An adaptive tuning procedure is proposed in [357] (see algorithm 2 therein) based on a Bayesian optimisation of the normalised expected jumping distance

$$J(\delta t, \tau/\delta t) \equiv \mathbb{E}\|z^{n+1} - z^n\|^2/\sqrt{\tau/\delta t},$$

where the expectation is taken with respect to the (HMC) transition kernel. Given a set of parameters, approximate evaluations of J over this set are formed by approximating the expectations using HMC. These evaluations are used to construct a Gaussian process surrogate for J, and finally a utility function is defined to propose suitable parameters. A good rate of convergence of the chain is also necessary for HMC to be of practical interest. If $\mathbb{P}_q^n$ denotes the conditional distribution of X_n given $X_1 = q$, we say the chain is **geometrically ergodic** if there exists $\varrho < 1$ and $C : \mathcal{M} \to [0, \infty)$ s.t. for P-a.e. $q \in \mathcal{M}$

$$\|\mathbb{P}_q^n - P\|_{TV} \leq C(q)\varrho^n.$$

Conditions under which the HMC chain with leapfrog integrator is geometrically ergodic were established in [235] for sufficiently small step-size under assumptions on the distribution of the number of leapfrog steps, the gradient of the potential energy (its growth and an inward-pointing condition), and the inward acceptance property (inward convergence of the acceptance region) of the chain. More recently [105] established geometric ergodicity under weaker assumptions, and also provided conditions on the potentials energy for which the inward acceptance property of the chain holds (see theorem 9-10). Geometric ergodicity is also shown in [55] for an idealised scenario in which the Hamiltonian flow is integrated exactly by setting the lengths of the trajectories to be i.i.d. exponential random variables.

A.4.6 Generative Models, and Likelihood-free Inference

Above we have mainly considered statistical models $\{P_\theta : \theta \in \Theta\}$ where the distributions were known, or which a likelihood was known up to normalisation $P_\theta \propto p_\theta \nu$. However sometimes the model is only known implicitely, in the sense that we only have at our disposal a procedure to generate samples from every P_θ. For example **generative models** are defined as the pushforward of a fixed probability measure $\mathbb{U}$ by a family of maps (G_θ) called the **generator**:

$$\{P_\theta \equiv G_\theta^\sharp \mathbb{U} : \theta \in \Theta\}.$$

Given a discrepancy function D on distributions, and assuming the target P belongs to the model, the aim is then to estimate

$$\theta^* \equiv \operatorname*{argmin}_{\theta \in \Theta} D(P \| G_\theta^\sharp \mathbb{U})$$

given a set (X_i) of independent P-distributed random variables, and a set (U_i) of independent $\mathbb{U}$-distributed random variables (so $G_\theta(U_i)$ are independent P_θ-distributed random variables).

We refer to 2.1 [270] for a summary of likelihood-free inference methods. A popular approach to estimate θ^* is the **adverserial framework**, where the discrepancy is the maxima among a family of functions (D_ϕ) called the **discriminator**, $D = \max_\phi D_\phi$, which leads us to the minimax problem

$$\theta^* \equiv \operatorname*{argmin}_{\theta \in \Theta} \max_\phi D_\phi(P \| G_\theta^\sharp \mathbb{U}).$$

The framework is "adversarial" because it is implemented as a two-player game, where we alternate between training the discriminator D_ϕ (over ϕ) to distinguish between the P and $G_\theta^\sharp \mathbb{U}$ (or more precisely their empirical version), and training the generator G_θ (over θ) to minimise the discrepancy between P and $G_\theta^\sharp \mathbb{U}$, i.e., to "confuse" the discriminator. In practice

each step is performed using a stochastic gradient method §5.2.

Originally [132] chose the discriminator to be

$$\begin{aligned} D_\phi(P\|G_\theta^\sharp\mathbb{U}) &\equiv \int \log M_\phi \mathrm{d}P + \int \log\left(1 - M_\phi\right) \mathrm{d}G_\theta^\sharp\mathbb{U} \\ &= \int \log M_\phi \mathrm{d}P + \int \log\left(1 - M_\phi \circ G_\theta\right) \mathrm{d}\mathbb{U} \end{aligned}$$

where both the generator and $M_\phi : \mathcal{M} \to [0,1]$ were multilayer perceptrons (M_ϕ is also often referred to as the discriminator). If P and P_θ do have positive densities p and p_θ respectively with respect to a common reference measure, then proposition/theorem 1 [132] show the best possible D_ϕ over the set of functions $M_\phi : \mathcal{M} \to [0,1]$ is given by setting $M^* \equiv p/(p + p_\theta)$, in which case it becomes

$$D^*(P\|G_\theta^\sharp\mathbb{U}) = 2\,\mathrm{JSD}(P\|G_\theta^\sharp\mathbb{U}) - \log 4$$

where JSD is the Jensen–Shannon statistical divergence, and thus θ^* becomes the minimum Jensen–Shannon discrepancy estimate.

In practice the **generative adversarial networks** (GANs) of [132] suffer from training instabilities resulting from the fact that training the discriminator often results in it attaining 0 (corresponding to maximal Jensen–Shannon discrepancy) due to the measures P_θ and P having disjoint support. This results in vanishing gradients that make it hard to then train the generator. In [280] GANs is generalised to allow D to be any f-divergence

$$D_f(P\|Q) \equiv \int p f\left(\frac{q}{p}\right) \mathrm{d}x$$

where p, q are the densities of P, Q, and $f : \mathbb{R}_+ \to \mathbb{R}$ is a convex lower-semicontinuous function with $f(1) = 0$. To cast the minimum f-divergence problem as an adversarial framework [280] relies on the variational characterisation of f-divergence: if $\mathcal{M} \subset \mathbb{R}^n$ is compact, and $\mathcal{F}$ is any set of functions $\mathcal{M} \to \mathbb{R}$, then (see lemma 1 [278])

$$D_f(Q\|P) \geq \sup_{h \in \mathcal{F}} \int \left(h\mathrm{d}Q - f^*(h)\mathrm{d}P\right)$$

with equality iff the map $x \mapsto \partial f\left(\frac{p}{q}(x)\right)$ contains an element of $\mathcal{F}$ (here f^* and $\partial f(t)$ are the conjugate dual function and subdifferential of f respectively). In particular if f is differentiable the supremum is attained by the function $f'(p/q)$. Table 1 [280] summarises some popular choices of f and their associated statistical divergences. Using the variational characterisation

yields the minimax problem

$$\theta^* \equiv \operatorname*{argmin}_\theta \max_h \int (h\mathrm{d}Q - f^*(h)\mathrm{d}P)\,.$$

See also Wasserstein GAN[16], MMD GAN [107], Sikhorn GAN [127], and Sobolev GAN [272].

A.4.7 Learning the Score

Often one is interested in specifically estimating the score $\nabla \log p$ of the target distribution $P(\mathrm{d}x) = p(x)\mathrm{d}x$, for example to implement a Langevin process or HMC algorithm [322, 330].

An option is to apply (2.9) with the score $\nabla \log q$ replaced with a family of differentiable maps $\{n_\theta : \mathbb{R}^n \to \mathbb{R}^n\}_{\theta\in\Theta}$ such as a neural network (here $\mathcal{M} = \mathbb{R}^n$), leading to the objective function (using $\Delta \log q = \nabla \cdot \nabla \log q = \operatorname{tr} \nabla^2 q$)

$$\theta \mapsto \int \left(\tfrac{1}{2}\|n_\theta\|^2 + \operatorname{tr} \nabla n_\theta\right) \mathrm{d}P.$$

Under regularity conditions, using the divergence theorem this objective is equivalent to $\theta \mapsto \int p\|n_\theta - \nabla \log p\|^2 \mathrm{d}x$ and (since $p > 0$) vanishes when $n_{\theta^*} = \nabla \log p$.

A related approach is based on autoencoders. An **autoencoder** is a neural network used in unsupervised learning to uncover some common structures of a set of observations. More precisely, let $\{x_i \in \mathbb{R}^n\}_i^N$ be a set of observations, $\xi_\theta : \mathbb{R}^n \to \mathbb{R}^m$ be an **encoder**, and $\delta_\alpha : \mathbb{R}^m \to \mathbb{R}^n$ be a **decoder** depending on parameters θ and α respectively, the autoencoder is trained to ensure the reconstruction $r \equiv \delta_\alpha \circ \xi_\theta$ is closed to the identity by finding the optimal reconstruction in the sense that

$$\operatorname*{argmin}_{\theta,\alpha} \frac{1}{N} \sum_i^N l(x_i, r(x_i)),$$

where l is a loss function penalising the discrepancy between $r(x)$ and x (such as a statistical divergence). The goal of course is not to actually learn the identity. By imposing appropriate constraints on the network, such as $m < n$, the encoder learns to "compress" the data to find out the "true" dimensions of the observations (removing redundancies, correlations...), while the decoder learns to recover the original observation given its compressed representation. In order to make the autoencoder robust to partial destruction of the observations, [354] proposed to modify the objective by introducing random variables $N_i \sim P(\cdot|x_i)$ distributed according to a regular conditional distribution $P(\cdot|\cdot)$ which corrupt the data points. We thus obtain the

denoising autoencoder (DAE) objective function

$$\operatorname*{argmin}_{\theta,\sigma} \frac{1}{N} \sum_{i}^{N} l(x_i, r(N_i)).$$

It was shown in 4 [353] that a DAE with sigmoid encoder, affine decoder, quadratic loss and Gaussian corruption was equivalent to a score matching estimation of the score of a Gaussian smoothing of the target with a particular statistical model. This was generalised in theorem 1 of [9] which shows a trained DAE with quadratic loss and corruption $N_i = x_i + \xi$, $\xi \sim \mathcal{N}(0, \sigma^2 \mathrm{Id})$ learns a reconstruction r_σ from which the score can be estimated when $\sigma \approx 0$:

$$\nabla \log p = \frac{r_\sigma(x) - x}{\sigma^2} + o(1).$$

This provides an alternative way of estimating the score of the target distribution by training a DAE avoiding the issues of normalising constants.

Deep MCMC [317] uses autoencoders to reduce the computational cost of HMC through dimensionality reduction. The idea is to train an encoder ξ and decoder δ using samples from the target P, to learn the "true" dimension of the samples. Once ξ and δ are trained, given current position q_0, each step samples a "momentum" $p_0 \sim \xi_\sharp \mathcal{N}(0, G)$, computes $(q^*, p^*) = \Psi(\xi(q_0), p_0)$, where Ψ is a numerical integrator for the Hamiltonian $\delta^* T + \delta^* V$, and outputs a sample $\delta(q^*)$. Note that the numerical integration is implemented on the lower dimensional space. However a bias is introduced unless a MC step is added, but since the proposal is not volume-preserving such an MC step would require computing an intractable volume correction term.

A.4.8 Sinkhorn Statistical divergences

As above consider $\mathcal{M} = \{x_1, \ldots, x_n\}$. To compute the Wasserstein distance (2.11) we need to find the optimal plan within the set $U(P, Q)$ of all transport plans between P and Q. [92] proposed to replace this by a convex set $U_\alpha(P, Q)$ obtained an entropic constraints

$$U_\alpha(P, Q) \equiv \{\boldsymbol{P}_{P \leftrightarrow Q} \in U(P, Q) : \mathrm{KL}\left(\boldsymbol{P}_{P \leftrightarrow Q} \| P \otimes Q\right) \leq \alpha\}$$

where KL here is the relative entropy between transport plans,

$$\mathrm{KL}(\boldsymbol{P} \| \boldsymbol{T}) \equiv \sum_{ij} \boldsymbol{P}_{ij} \log \frac{\boldsymbol{P}_{ij}}{\boldsymbol{T}_{ij}}.$$

The regularised Wasserstein or **Sinkhorn discrepancy** is defined as

$$W_1^\alpha(P,Q) \equiv \min_{\boldsymbol{P}_{P\leftrightarrow Q}\in U_\alpha(P,Q)} \boldsymbol{P}^\top D.$$

Equivalently, replacing the entropy constraint by a regularisation term parametrised by $\epsilon > 0$ we can consider

$$W_{1,\epsilon}(P,Q) \equiv \min_{\boldsymbol{P}_{P\leftrightarrow Q}\in U(P,Q)} \boldsymbol{P}^\top D + \epsilon\,\mathrm{KL}\left(\boldsymbol{P}\|P\otimes Q\right). \tag{A.5}$$

Setting $(K_{ij}) \equiv \left(e^{-D_{ij}/\epsilon}\right)$, by proposition 2.1 [47] there exists nonnegative vectors u, v s.t., the optimal transport plan is

$$\boldsymbol{P}^*_{P\leftrightarrow Q} = \mathrm{diag}(u) K\,\mathrm{diag}(v);$$

moreover these vectors are fixed point of the map $(u,v) \mapsto \left(P/(Kv), Q/(K^\top u)\right)$, where $/$ denotes component-wise division. **Sinkhorn's algorithm** for the solutions (u,v) is defined by solving these equations iteratively: given an arbitrary initial positive vector $v^{(0)}$, at step $k+1$ the algorithm perform the following updatev(see 4.2 [298])

$$u^{(k+1)} \equiv P/(Kv^{(k)}), \qquad v^{(k+1)} \equiv Q/(K^\top u^{(k+1)}).$$

The Sinkhorn discrepancy $W_{1,\epsilon}$ is not a natural map to represent discrepancies between distributions, since for example $W_{1,\epsilon}(P,P)$ is not zero. Instead we can define the **Sinkhorn statistical divergence**

$$\mathcal{S}_{1,\epsilon}(P,Q) \equiv W_{1,\epsilon}(P,Q) - \tfrac{1}{2}\left(W_{1,\epsilon}(P,P) + W_{1,\epsilon}(Q,Q)\right).$$

Then

$$\mathcal{S}_{1,\epsilon}(P,Q) \xrightarrow{\epsilon\to 0} W_1(P,Q), \qquad \mathcal{S}_{1,\epsilon}(P,Q) \xrightarrow{\epsilon\to+\infty} \tfrac{1}{2}\,\mathrm{MMD}_d(P,Q)$$

where MMD_d is the maximum mean discrepancy with kernel d. Moreover under appropriate assumption $\mathcal{S}_{1,\epsilon}$ is a symmetric statistical divergence that controls narrow convergence (see theorem 1 [119]).

A.4.9 Exponential Family

The exponential family is the parametrised statistical manifold defined by the likelihood

$$p_\theta(x) \equiv \frac{1}{Z(\theta)} e^{\langle v(\theta), T(x)\rangle_*}$$

with respect to some reference measure $\nu(\mathrm{d}x)$, where $v : \Theta \to V$, $T : \Omega \to V^*$ and V is a topological vector space. For example if $V \equiv \mathcal{H}$ is a Hilbert space of functions over Ω, then

$T : \Omega \to \mathcal{H}$ could be the evaluation functionals $x \mapsto \delta_x$, $\Theta = \{f \in \mathcal{H} : \int e^{f(x)}\nu(\mathrm{d}x) < \infty\}$ the subset of $\mathcal{H}$ for which the normalising constant is finite, $Z(\theta) < \infty$, and v the inclusion, so

$$p_\theta(x) \equiv \frac{1}{Z(\theta)} e^{\langle \theta, \delta_x \rangle_*} = \frac{1}{Z(\theta)} e^{\theta(x)}.$$

In particular when $\mathcal{H}$ is a Reproducing Kernel Hilbert space with kernel $k : \Omega \times \Omega \to \mathbb{R}$, we have $\theta(x) = \langle \theta, \delta_x \rangle_* = \langle \theta, k(x, \cdot) \rangle_{\mathcal{H}}$ by the reproducing property. For finite dimensional exponential families, $V \equiv \mathbb{R}^d$, we set $T = (f_1, \ldots, f_d)$, $\Theta \equiv \{z \in \mathbb{R}^d : \int e^{z^\top T(x)} < \infty\}$, so

$$p_\theta(x) \equiv \frac{1}{Z(\theta)} e^{\langle \theta, T(x) \rangle_*} = \frac{1}{Z(\theta)} e^{T_i(x)\theta^i}.$$

In this case the components of the score are $\partial_{\theta^i} \log p_\theta(x) = T_i(x) - \frac{\partial}{\partial \theta^i} \log Z(\theta)$, and the Fisher information matrix is

$$\left((\phi^* g^F(\theta)\right)_{ij} = \frac{\partial^2}{\partial \theta^i \partial \theta^j} \log Z(\theta) = -\frac{\partial^2}{\partial \theta^i \partial \theta^j} \log p_\theta(x)$$

A.4.10 Discrete Space Diffusion and Markov Chains

In §4.1 we have focused on P-preserving stochastic processes which were continuous in time and (state) space, that is $Z : \mathbb{R} \times \Omega \to \mathbb{R}^n$, $Z(t, \cdot) \equiv Z_t(\cdot)$. The discretisation in §4.1.2 gave an example of discrete-time, continuous space $Z : \mathbb{N} \times \Omega \to \mathbb{R}^n$. It is also possible to consider discrete Markov chains in which both time and space are countable, $Z : \mathbb{N} \times \Omega \to \mathbb{Z}$. For time-homogeneous Markov chains the transition kernel is usually indexed by state space,

$$\mathbb{P}_{ij} \equiv \mathbb{P}(X_m = i \to X_{m+1} = j), \qquad \mathbb{P}^n_{ij} \equiv \mathbb{P}(X_m = i \to X_{m+n} = j).$$

The Chapman–Kolmogorov equation states the $(n+m)$-step transition kernel is related to the n and m steps ones through a "matrix product"

$$\mathbb{P}^{n+m}_{ij} = \mathbb{P}^n_{ik} \mathbb{P}^m_{kj} \qquad \Rightarrow \qquad (\mathbb{P}^n_{ij}) = (\mathbb{P}_{sk})^n,$$

from which it follows that the law P_{X_k} of X_k, whose i^{th} component is the probability that $X_k = i$, is determined by the distribution of the initial point P_{X_0} (see 2.1 [294])

$$P_{X_k} = P_{X_0} \mathbb{P}^k, \qquad \text{or} \qquad (P_{X_k})_j = (P_{X_0})_i \, \mathbb{P}^k_{ij}.$$

A.5 Further Material HMC

A.5.1 Riemannian Measure

Any Riemannian manifold $(\mathcal{M}, \langle\cdot,\cdot\rangle^{\mathcal{M}})$ has a Radon measure compatible with the Riemannian structure. The measure is constructed by first defining the Riemannian one density vol, which maps any orthonormal basis of $T_q\mathcal{M}$ to one (or to its orientation on oriented manifold, in which case vol is a volume form), and then use Riesz representation to turn vol into a measure $\mathrm{d}V$ s.t., $\int f\mathrm{d}V \equiv \int f\mathrm{vol}$ (theorem 8.1.9 [3]). If (v_i) is a local basis of vector fields over a neighbourhood $U \subset \mathcal{M}$, with dual local 1-forms (α^i), then

$$\mathrm{vol}|_U = \sqrt{\det(G)}|\alpha^i \wedge \cdots \wedge \alpha^m|$$

where $\det(G) \equiv \det\left(G(v_i, v_j)\right)$. If q^i are coordinates on $\mathcal{M}$ generated by a chart $\varphi : U \to \mathbb{R}^n$, we may for example choose $v_i = \partial_{q^i}$, $\alpha^i = \mathrm{d}q^i$, we can write the integral of any $f \in L^1(\mathcal{M}, \mathrm{vol})$ with compact support contained in U as an integral over $\mathbb{R}^n$

$$\int_{\mathcal{M}} f\mathrm{d}V = \int_{\varphi(U)} f \circ \varphi^{-1}\sqrt{\det G} \circ \varphi^{-1}\mathrm{d}\lambda^n,$$

where $\mathrm{d}\lambda^n$ is the Lebesgue measure on $\mathbb{R}^n$. In particular if $\iota : \mathcal{M} \to \mathbb{R}^k$ is a submanifold, then $\mathcal{M}$ can be equipped with the pullback (or restriction) of the Euclidean dot product $\langle\cdot,\cdot\rangle^{\mathcal{M}} \equiv \iota^* \langle\cdot,\cdot\rangle^{\mathbb{R}^k} = \partial\iota(\cdot)^\top M \partial\iota(\cdot)$, where M is an inner product matrix on $\mathbb{R}^k$. If q^i are coordinates over the chart $\varphi : U \to \mathbb{R}^n$

$$\int_{\mathcal{M}} f\mathrm{vol}_g = \int_{\varphi(U)} f \circ \varphi^{-1}\sqrt{\det\left(\partial J^\top M \partial J\right)}\mathrm{d}\lambda^n,$$

where ∂J is the Jacobian of $J \equiv \iota \circ \varphi^{-1} : \varphi(U) \subset \mathbb{R}^n \to \mathbb{R}^k$ (this follows from $\langle\cdot,\cdot\rangle^{\mathcal{M}}_{ij} \equiv \left\langle \partial_{q^i}, \partial_{q^j} \right\rangle^{\mathcal{M}} = \left\langle \partial\iota(\partial_{q^i}), \partial\iota(\partial_{q^j}) \right\rangle^{\mathbb{R}^k} = (\partial J^\top M \partial J)_{ij}$). The term $\sqrt{\det\left(\partial J^\top M \partial J\right)}\mathrm{d}\lambda^n$ is sometimes referred to as the **volume element** of $\mathcal{M}$ since $\sqrt{\det \partial J^\top M \partial J}$ is the volume of the parallelepiped with corners $(\partial_i J)$ which describe the deformation of "infinitesimal volume elements" on $\mathcal{M}$ with respect to Cartesian coordinates. Setting $f = 1$ yields the volume of U in $\mathcal{M}$, which may also be expressed as the Hausdorff measure $\mathcal{H}^n(U)$ of U (see 3.3.4 D [110]). We mention that in the statistics literature, the Riemannian measure is sometimes introduced as a **Hausdorff measure**. Given a measure μ on $\mathcal{M}$ and a finite measurable set K, the **uniform density** π_K with respect to μ on K is defined as $\pi_K(x) \equiv 1_K(x)/\mu(K)$. This definition depends

on the reference measure μ and should not be confused with the canonical concept of uniform measure discussed in §3.1.2 (see also 2.1 in [295]).

A.5.2 Retractions

When the geodesic flow has a prohibitive computational cost, we can consider replacing the flow of the kinetic energy with a first-order approximation, known as a **retraction** [5]. To be precise a retraction is a map $R : T\mathcal{M} \to \mathcal{M}$ s.t. $R_q \equiv R|_{T_q\mathcal{M}}$ [12]

$$\textbf{(i)} \quad R_q(0) = q, \qquad \textbf{(ii)} \quad \partial_0 R_q = \mathrm{Id}_{T_q\mathcal{M}}.$$

Retractions are first-order accurate in the sense that

$$d\left(R_q(tv), \Phi_t^T(q,v)\right) = O(t^2), \qquad \forall v \in T_q\mathcal{M}$$

where $d : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ is the Riemannian distance (see proposition 2.2 [6]). For example for the B-orthogonal group $\mathcal{G} = \{A : \mathrm{GL}(n) : A^T B A = B\}$ for some matrix $B \in \mathrm{GL}(n)$ (which include the orthogonal and symplectic groups), the Cayley transform $R : \mathcal{G} \times \mathfrak{g} \to \mathcal{G}$ given by $R_A(\xi) \equiv A(I - \frac{1}{2}\xi)^{-1}(I + \frac{1}{2}\xi)$ approximates the Lie group exponential Ae^{ξ} and stays exactly on $\mathcal{G}$ [75, 147]. It is possible to build retractions on $\mathcal{G}/\mathcal{K}$ using retractions on $\mathcal{G}$, see proposition 4.1.3 of [5], and the following page for examples on the real projective space and Grassmannian manifold. On general manifolds another method to build retractions is to use an embedding into a vector space, as explained in proposition 4.1.2 [5], or theorem 4.2 [6]. This idea was for example applied in [363] for HMC on Stiefel manifolds, while Cayley transforms are sometimes used in the implementation of HMC for lattice QCD.

A.5.3 Principle of Stationary Action

Consider the smooth manifold $\mathcal{C}$ of C^2 curves $\gamma : [0,T] \to \mathcal{M}$. Given an interval $I \subset \mathbb{R}$, a **variation** $\sigma : I \times [0,T] \to \mathcal{M}$ of γ is a C^2 map satisfying $\sigma(0,\cdot) = \gamma$, and we say it has **fixed endpoints** if

$$\sigma(\cdot, 0) = \gamma(0), \qquad \sigma(\cdot, T) = \gamma(T).$$

Define the **variation field** $\delta\sigma : [0,T] \to T\mathcal{M}$ along γ by

$$\delta\sigma(t) \equiv \frac{\mathrm{d}}{\mathrm{d}s}\sigma(s,t)\Big|_{s=0} \in T_{\sigma(0,t)}\mathcal{M}.$$

[12] Where we identify $T_0 T_q\mathcal{M} \cong T_q\mathcal{M}$ so $\partial_0 R_q : T_q\mathcal{M} \to T_q\mathcal{M}$.

We may then consider variations of the action with respect to a variation field:

$$\begin{aligned}\delta\mathcal{S}[\sigma(s,\cdot)] \equiv \frac{\mathrm{d}}{\mathrm{d}s}\mathcal{S}[\sigma(s,\cdot)]\big|_{s=0} &= \frac{\mathrm{d}}{\mathrm{d}s}\int_0^T \mathcal{L}\left(\dot{\sigma}(s,\cdot)\right)\mathrm{d}t\big|_{s=0} \\ &= \int_0^T \frac{\mathrm{d}}{\mathrm{d}s}\mathcal{L}\left(\dot{\sigma}(s,\cdot)\right)\big|_{s=0}\mathrm{d}t = \int_0^T \mathrm{d}\mathcal{L}\left(\delta\dot{\sigma}(t)\right)\mathrm{d}t \\ &= \langle \mathbb{F}\mathcal{L}^*\Theta, \delta\sigma\rangle_*\,\big|_0^T + \int_0^T \langle D_{EL}\mathcal{L}, \delta\sigma\rangle_*\,\mathrm{d}t\end{aligned}$$

where $D_{EL}\mathcal{L}$ is a continuous $T^*\mathcal{M}$-valued map defined on an appropriate submanifold of $TT\mathcal{M}$ (see [252] for details on last step), called the Euler-Lagrange map since in coordinates

$$(D_{EL}\mathcal{L})_i = \frac{\partial\mathcal{L}}{\partial q^i} - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial\mathcal{L}}{\partial v^i}\right).$$

We say a curve $\gamma \in \mathcal{C}$ is a solution of **Euler-Lagrange equations** if

$$\frac{\partial\mathcal{L}}{\partial q^i}(q,\dot{q}) - \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial\mathcal{L}}{\partial v^i}(q,\dot{q})\right) = 0$$

for all $t \in (0,T)$, or equivalently $\int_0^T \langle D_{EL}\mathcal{L}, \delta\sigma\rangle_*\,\mathrm{d}t = 0$ for all variations of γ, or equivalently if $\delta\mathcal{S}[\sigma] = 0$ for all variations with fixed endpoints, or equivalently if $\dot{\gamma}$ is an integral curve of the Lagrangian vector field defined below (see 8.2 and theorem 7.3.3 [251]).

A.5.4 Motion in a Potential Field Vs Free Motion

If the potential energy is bounded above by some $e \in \mathbb{R}$, $V < e$, we can define the **Jacobi metric** as[13]

$$G_e(\cdot,\cdot) \equiv (e - V)\langle\cdot,\cdot\rangle,$$

which is a Riemannian metric on $\mathcal{M}$. The importance of the Jacobi metric is that it relates motion under the influence of the potential field V to free motion. More precisely, the solution of the Euler-Lagrange equation for $\mathcal{L}(\cdot) \equiv \frac{1}{2}\langle\cdot,\cdot\rangle - \pi^*V(\cdot)$ with energy $E = e$ are (up to reparametrisation) the geodesics of the Jacobi metric G_e with energy 1 (with respect to the Lagrangian defined by G_e) [251]. Since the Lagrangian used in HMC are usually of this form, the Jacobi metric allows us to reformulate some statistical questions in terms of properties of the geodesics. For example, the Jacobi metric was used in [311] to derive a concentration inequality for HMC that quantifies how many HMC samples are necessary for the empirical measure to provide a good approximation of the target distribution, in the case the target

[13] In other words G_e and $\langle\cdot,\cdot\rangle$ are conformally related by the function $\varrho \equiv \log\sqrt{e-V}$, since $G_e = e^{2\varrho}\langle\cdot,\cdot\rangle$

distribution is a multivariate Gaussian on $\mathcal{M} = \mathbb{R}^n$ and the chain has positive curvature. This was achieved by deriving a bound between the coarse Ricci curvature of the HMC chain targeting the multivariate Gaussian and the sectional curvature of $\mathbb{R}^n$ equipped with the Jacobi metric in high dimensions.

A.5.5 Discrete Variational Principle

The main point in the construction of the discrete variational principle is to replace the velocity space $T\mathcal{M}$ with the product $\mathcal{M} \times \mathcal{M}$, and the Lagrangian $\mathcal{L} : T\mathcal{M} \to \mathbb{R}$ with a family of **discrete Lagrangians** $\mathcal{L}_d : \mathcal{M} \times \mathcal{M} \times \mathbb{R} \to \mathbb{R}$ parametrised by step size $\delta t \in \mathbb{R}$. The discrete Lagrangian represents an approximation of the "infinitesimal action" of a curve $q : [t, t + \delta t] \to \mathcal{M}$

$$\mathcal{L}_d(q(t), q(t + \delta t), \delta t) \approx \int_t^{t+\delta t} \mathcal{L}(\dot{q})\mathrm{d}t.$$

Given a step size δt and a $N \in \mathbb{N}$, the manifold of C^2 curves $[0, T] \to \mathcal{M}$ is replaced by the product $\mathcal{M}^{N+1}$ where $(q_0, \ldots, q_N) \in \mathcal{M}^{N+1}$ represents the values of a curve evaluated at the points $\{j\delta t : j = 0, 1, \ldots, N\}$. Forming the **discrete action** $\mathcal{S}_d : \mathcal{M}^{N+1} \to \mathbb{R}$

$$\mathcal{S}_d(q_0, \ldots, q_N) \equiv \sum_{j=0}^{j=N-1} \mathcal{L}_d(q_j, q_{j+1}, \delta t)$$

and taking variations, leads us to the **discrete Euler–Lagrange equations** whose solutions are discrete curves $(q_0, \ldots, q_N) \in \mathcal{M}^{N+1}$ satisfying in local coordinates

$$\partial_2 \mathcal{L}_d(q_{j-1}, q_j, \delta t) + \partial_1 \mathcal{L}_d(q_j, q_{j+1}, \delta t) = 0, \qquad \text{for all } j = 1, \ldots, N - 1. \tag{A.6}$$

The discrete Euler–Lagrange equations define a **discrete Lagrangian flow** $\Phi_{\mathcal{L}_d} : \mathcal{M} \times \mathcal{M} \to \mathcal{M} \times \mathcal{M}$, $(q_{j-1}, q_j) \mapsto (q_j, q_{j+1})$ which preserves a discrete symplectic structure (see 1.3.2 [252]).

We can define a corresponding conjugate momentum and discrete Hamiltonian flow by considering the discrete Legendre transform $\mathbb{F}\mathcal{L}_d : \mathcal{M} \times \mathcal{M} \to T^*\mathcal{M}$ defined by

$$(q_j, p_j) \equiv \mathbb{F}\mathcal{L}_d(q_j, q_{j+1}) \equiv (q_j, -\partial_1 \mathcal{L}_d(q_j, q_{j+1}, \delta t)) .$$

Assuming $\mathbb{F}\mathcal{L}_d$ is a local isomorphism, theorem 1.5.1 [252] implies the discrete Lagrangian flow $\Phi_{\mathcal{L}_d}$ is well-defined, and we can define the **discrete Hamiltonian flow** $\Phi^H_{\mathcal{L}_d} : T^*\mathcal{M} \to T^*\mathcal{M}$ by

$$\Phi^H_{\mathcal{L}_d} \equiv \mathbb{F}\mathcal{L}_d \circ \Phi_{\mathcal{L}_d} \circ \mathbb{F}\mathcal{L}_d^{-1}.$$

This map is symplectic, and in local coordinates the discrete Hamiltonian flow is the map $(q_j, p_j) \mapsto (q_{j+1}, p_{j+1})$ defined by

$$p_j = -\partial_1 \mathcal{L}_d(q_j, q_{j+1}, \delta t), \qquad p_{j+1} = \partial_2 \mathcal{L}_d(q_j, q_{j+1}, \delta t). \tag{A.7}$$

In order for variational integrators to be applicable to HMC they will need to also be time-reversible and symmetric (see §3.4.1). These may be ensured using the following results. The discrete analogue to the velocity flip is the map $S_d(q_0, q_1) \equiv (q_1, q_0)$, and we say that $\mathcal{L}_d$ is reversible (or has time-reversal symmetry) if $\mathcal{L}_d \circ S_d = \mathcal{L}_d$. In that case the discrete Lagrangian flow is **discrete time-reversible** (2.2 and 3.3 [257])

$$\Phi_{\mathcal{L}_d} \circ S_d \circ \Phi_{\mathcal{L}_d} = S_d.$$

Moreover $\mathcal{L}_d(q_0, q_1, -\delta t) = -\mathcal{L}_d(q_1, q_0, \delta t)$ iff the discrete Hamiltonian flow is symmetric (theorem 2.4.1 [252]). Finally, if the discrete Lagrangian inherits the symmetries of $\mathcal{L}$, then the the discrete flows $\Phi_{\mathcal{L}_d}$, $\Phi^H_{\mathcal{L}_d}$ will also preserve these symmetries (1.5.3 and theorem 1.3.3. [252]). In what follows we briefly mention some examples of variational integrators, in particular the generalised leapfrog on which Riemannian "manifold" HMC is built [131].

A.5.5.1 Generalised Leapfrog

Given a Lagrangian $\mathcal{L}$ on $\mathcal{M} = \mathbb{R}^n$, the trapezoidal rule defines a discrete Lagrangian

$$\mathcal{L}_d(q_k, q_{k+1}, \delta t) \equiv \frac{\delta t}{2}\mathcal{L}\left(q_k, \frac{q_{k+1} - q_k}{\delta t}\right) + \frac{\delta t}{2}\mathcal{L}\left(q_{k+1}, \frac{q_{k+1} - q_k}{\delta t}\right).$$

For $\mathcal{L} \equiv \frac{1}{2}v^\top M v - V(q)$ this yields the discrete Lagrangian flow

$$q_{k+1} = 2q_k - q_{k-1} - \delta t^2 M^{-1} \nabla V(q_k)$$

and discrete Hamiltonian flow with

$$\begin{aligned} q_{k+1} &= q_k + \delta t M^{-1} p_k - \tfrac{1}{2}\delta t^2 M^{-1} \nabla V(q_k) \\ p_{k+1} &= p_k - \delta t \frac{\nabla V(q_k) + \nabla V(q_{k+1})}{2} \end{aligned}$$

which are also known as Verlet and velocity Verlet respectively. In the HMC literature the latter is usually written as

$$p_{k+\frac{1}{2}} = p_k - \tfrac{1}{2}\delta t \nabla V(q_k)$$

$$q_{k+1} = q_k + \delta t M^{-1} p_{k+\frac{1}{2}}$$
$$p_{k+1} = p_{k+\frac{1}{2}} - \tfrac{1}{2}\delta t \nabla V(q_{k+1}).$$

For a space-dependent metric, $\mathcal{L} \equiv \frac{1}{2} v^\top M(q) v - V(q)$, (A.7) yields the **generalised leapfrog method** which is used in RMHMC [131]

$$p_j = \tfrac{1}{2}\left(M(q_j) + M(q_{j+1})\right)\delta q - \frac{\delta t}{2}\nabla_{q_j}\left(\delta q^\top M(q_j)\delta q - V(q_j)\right)$$
$$p_{j+1} = \tfrac{1}{2}\left(M(q_j) + M(q_{j+1})\right)\delta q + \frac{\delta t}{2}\nabla_{q_{j+1}}\left(\delta q^\top M(q_{j+1})\delta q - V(q_{j+1})\right)$$

where $\delta q \equiv \frac{q_{j+1}-q_j}{\delta t}$. Here the first equation defines q_{j+1} implicitly, which is then subbed into the second (explicit) equation for p_{j+1}.

A.5.5.2 Midpoint Rule

For a Hamiltonian H, the midpoint rule is the symplectic integrator implicitely defined by

$$\frac{q_{k+1} - q_k}{\delta t} = \frac{\partial H}{\partial p}\left(\frac{q_{k+1} + q_k}{2}, \frac{p_{k+1} - p_k}{2}\right)$$
$$\frac{p_{k+1} - p_k}{\delta t} = -\frac{\partial H}{\partial q}\left(\frac{q_{k+1} + q_k}{2}, \frac{p_{k+1} - p_k}{2}\right)$$

can be obtained (if $\mathcal{L}$ is regular) from the discrete Lagrangian (see 2.6.1 [252])

$$\mathcal{L}_d(q_k, q_{k+1}, \delta t) \equiv \delta t \mathcal{L}\left(\frac{q_k + q_{k+1}}{2}, \frac{q_{k+1} - q_k}{\delta t}\right)$$

A.5.5.3 Symplectic partitioned Runge–Kutta methods

Any symplectic partitioned Runge–Kutta methods may be obtained from a discrete Lagrangian as shown in 2.6.5 [252].

A.5.6 Hamilton-Jacobi PDE

We now discuss the Hamilton–Jacobi PDE. In §A.5.6.1 we will introduce the extended phase space, a contact manifold, define the notion of time-dependent Hamiltonian vector fields in §A.5.6.2, and introduce canonical transformation of time-dependent fields in §A.5.6.3. With these definitions we can then rigorously formulate the Hamilton-Jacobi PDE which describes the time evolution of the generating function of a Hamiltonian flow; this will be done §A.5.6.4.

A.5.6.1 Extended Phase Space and Contact Manifolds

We now consider time-dependent mechanical systems, in which the Hamiltonian may depend on time t and is a function on the extended phase $H : T^*\mathcal{M} \times \mathbb{R} \to \mathbb{R}$. The extended phase space cannot be a symplectic manifold due to its $(2n+1)$-dimensionality. By proposition 5.1.9 [2], it is instead an (exact) **contact manifold** with **contact form** $\lambda \in \Gamma(T^*(T^*\mathcal{M} \times \mathbb{R}))$

$$\lambda \equiv \mathrm{d}t + \pi^*_{T^*\mathcal{M}}\Theta \tag{A.8}$$

which means that $\lambda \wedge (\mathrm{d}\lambda)^n$ is a volume form on $T^*\mathcal{M} \times \mathbb{R}$. [14] Such contact systems were used in [45] to define the adiabatic Monte Carlo algorithm.

A.5.6.2 Time-dependent Hamiltonian Flows

The **time-dependent Hamiltonian vector field** $X_H : T^*\mathcal{M} \times \mathbb{R} \to TT^*\mathcal{M}$ of H is defined to be for each $t \in \mathbb{R}$ the Hamiltonian vector field of $H_t \equiv H(\cdot, t)$, that is $((q,p),t) \mapsto X_{H_t}((q,p))$. Note that X_H is not a vector field! The description of time-dependent system is instead based on its **suspension** $\tilde{X}_H \in \Gamma(T(T^*\mathcal{M} \times \mathbb{R})) \cong \Gamma(TT^*\mathcal{M} \times T\mathbb{R})$, $\tilde{X}_H((q,p),t) = \frac{\partial}{\partial t} + X_H$. On a canonical coordinate system for $T^*\mathcal{M}$ we have

$$\tilde{X}_H((q,p),t) = \frac{\partial}{\partial t} + X_H((q,p),t) = \frac{\partial}{\partial t} + \frac{\partial H_t}{\partial p_i}\frac{\partial}{\partial q^i} - \frac{\partial H_t}{\partial q^i}\frac{\partial}{\partial p_i}.$$

Recall that a time-independent Hamiltonian vector field conserves energy and the symplectic volume form. As may be expected, energy is no longer preserved along the suspension: indeed from the coordinate expression we have

$$\mathcal{L}_{\tilde{X}_H} H = \frac{\partial H}{\partial t}.$$

However the suspension preserves the extended (or Poincaré–Cartan) forms

$$\Theta_H \equiv \pi^*_{T^*\mathcal{M}}\Theta - H\mathrm{d}t, \qquad \omega_H \equiv -\mathrm{d}\Theta_H = \pi^*_{T^*\mathcal{M}}\omega + \mathrm{d}H \wedge \mathrm{d}t \tag{A.9}$$

which means (see theorem 5.1.14 [2])

$$\mathcal{L}_{\tilde{X}_H}\omega_H = 0, \qquad \mathrm{d}\mathcal{L}_{\tilde{X}_H}\theta_H = 0,$$

[14] Note here that technically we view the time coordinate t as a function $T^*\mathcal{M} \times \mathbb{R} \to \mathbb{R}$ (which projects onto the $\mathbb{R}$-component). Also $\pi_{T^*\mathcal{M}} : T^*\mathcal{M} \times \mathbb{R} \to T^*\mathcal{M}$ is the usual projection.

and it preserves the volume form $\mathrm{d}t \wedge \omega_H^n = \mathrm{d}t \wedge (\pi_{T^*\mathcal{M}}^* \omega)^n$. The extended two-form provides an alternative definition of $\tilde{X}_H$: it is the unique vector field satisfying

$$i_{\tilde{X}_H} \omega_H = 0, \qquad \mathrm{d}t(\tilde{X}_H) = 1$$

A.5.6.3 Generating Functions of Canonical Transformations

A **canonical transformation** is one that preserves the equations of motion. In the time-independent case, we say that a diffeomorphism F is canonical if it preserves the symplectic structures (that is F is a symplectomorphism), since this is equivalent to the invariance of Hamilton's equations under F in the sense that $X_H = F_* X_{F^*H}$ (theorem 3.3.19 [2]).

In the time-dependent case, a time-preserving (i.e., $F^*t = t$) diffeomorphism $F : T^*\mathcal{M} \times \mathbb{R} \to T^*\mathcal{M} \times \mathbb{R}$ is called canonical if there exists a function $K_F : T^*\mathcal{M} \times \mathbb{R} \to \mathbb{R}$ s.t. $F^*\pi_{T^*\mathcal{M}}^*\Theta - \pi_{T^*\mathcal{M}}^*\Theta + K_F \mathrm{d}t$ is closed, or equivalently there locally exists a function $S : T^*\mathcal{M} \times \mathbb{R} \to \mathbb{R}$ s.t.

$$F^*\pi_{T^*\mathcal{M}}^*\Theta - \pi_{T^*\mathcal{M}}^*\Theta + K_F \mathrm{d}t = \mathrm{d}S. \tag{A.10}$$

By theorem 5.2.14 [2] this is equivalent to the invariance of the time-dependent Hamilton's equations: for any $H : T^*\mathcal{M} \times \mathbb{R} \to \mathbb{R}$

$$\tilde{X}_H = F_* \tilde{X}_{F^*H + K_F}.$$

The intuition behind this definition is provided by corollary 5.2.11 [2], which shows that the existence **generating function** S as above (which is equivalent to F being canonical) implies that $\pi_1 \circ F(\cdot, t) : T^*\mathcal{M} \to T^*\mathcal{M}$ is symplectic for any t.

A.5.6.4 Hamilton-Jacobi Equations and Jacobi's Solution

Recall from §3.5.3 that in the time-independent case we had that a map $F : T^*\mathcal{M} \to T^*\mathcal{M}$ was symplectic iff it had a generating function. In particular solutions to the generating equations (3.13) defined a symplectic map $F : (q, p) \mapsto (Q, P)$. Suppose that the symplectic map F was in fact a Hamiltonian flow $F \equiv \Phi_t^H$ at a given t. We want to now investigate how the generating function for Φ_t^H varies with t, i.e., we are interested in the map $t \mapsto S_t$ where S_t is the generating function for Φ_t^H, which solves for each t the generating equations

$$p = -\frac{\partial S_t}{\partial q}(q, Q(t)), \qquad P(t) = \frac{\partial S_t}{\partial Q}(q, Q(t)).$$

where $(Q(t), P(t)) \equiv \Phi_t^H(q,p)$ (note that for $t=0$ we should be generating the identity transformation). This will be the content of Hamilton-Jacobi PDE.

To do so we consider we lift $\Phi^H : T^*\mathcal{M} \times \mathbb{R} \to T^*\mathcal{M}$ (recall that here H is time-independent) to the transformation of the extended phase space $F : T^*\mathcal{M} \times \mathbb{R} \to T^*\mathcal{M} \times \mathbb{R}$, $F(z,t) \equiv (\Phi_t^H(z), t)$. Then for the "time-dependent" Hamiltonian $\pi_{T^*\mathcal{M}}^* H : T^*\mathcal{M} \times \mathbb{R} \to \mathbb{R}$, we have

$$F^*\pi_{T^*\mathcal{M}}^* H = H \circ \pi_{T^*\mathcal{M}} \circ F = H \circ \Phi^H$$

and from theorem 5.2.19 [2] F is canonical and $K_F = -H \circ \Phi^H$. From §A.5.6.3 we thus obtain

$$\tilde{X}_{\pi_{T^*\mathcal{M}}^* H} = F_* \tilde{X}_0 = F_* \frac{\partial}{\partial t},$$

and we say that F has transformed H to equilibrium. Viewing $S(t,q,Q)$ as a function of time and the coordinates (q,Q) on the graph of $\Gamma(\Phi^H) \subset T^*\mathcal{M} \times \mathbb{R} \times T^*\mathcal{M}$, we have that F transforms H to equilibrium iff S is a solution of the **Hamilton-Jacobi PDE** (page 390 [2])

$$H\left(Q, \frac{\partial S}{\partial Q}\right) = -\frac{\partial S}{\partial t}. \tag{A.11}$$

In fact since $\pi_{T^*\mathcal{M}} \circ F = \Phi^H$ and F is canonical, we know from (A.10) that locally

$$\Theta \circ \Phi^H - \pi_{T^*\mathcal{M}}^* \Theta - H \circ \Phi^H \mathrm{d}t = \mathrm{d}S$$

or in coordinates

$$P(t)\mathrm{d}Q(t) - p\mathrm{d}q - H(Q(t), P(t))\mathrm{d}t = \frac{\partial S}{\partial q}\mathrm{d}q + \frac{\partial S}{\partial Q}\mathrm{d}Q + \frac{\partial S}{\partial t}\mathrm{d}t$$

which comparing coefficients yields

$$\begin{aligned} p &= -\frac{\partial S}{\partial q}(t,q,Q(t)) \\ P(t) &= \frac{\partial S}{\partial Q}(t,q,Q(t)) \\ H\left(Q(t), \frac{\partial S}{\partial Q}(t,q,Q(t))\right) &= -\frac{\partial S}{\partial t}(t,q,Q(t)). \end{aligned}$$

Note here that the first two equations are simply the generating equations for Φ_t^H at time t, while the Hamilton-Jacobi PDE characterises how S varies in time. [15]

[15] By **Jacobi's solution** the solution of the Hamilton-Jacobi PDE (for the Hamiltonian of a regular La-

A.5.6.5 Connections with Mechanics

Taking the partial exterior derivative of a divergence defines a map $\mathrm{d}_1 D : \Theta \times \Theta \to T^*\Theta$. It was shown in [25] that the induced 2-form

$$\omega_D \equiv (\mathrm{d}_1 D)^* \, \omega$$

is a symplectic structure (at least on a neighbourhood of the diagonal $\Theta \times \Theta$). In local coordinates

$$\omega_D(x, y) = \frac{\partial^2 D(x\|y)}{\partial x^i \partial y^j} \mathrm{d}x^i \wedge \mathrm{d}y^j.$$

Having a symplectic structure enables us to lower/raise tensor indices, as well as define a volume form adapted to the information divergence. As was noted in [216] this is precisely the discrete symplectic form induced by the divergence D viewed as a discrete Lagrangian on Θ. The paper also proposes using the so-called canonical divergence to measure the **duality gap**, that is the discrepancy between a Hamiltonian system and a Lagrangian system.

A.5.7 Additional Proofs HMC

A.5.7.1 Magnetic HMC

Derivation Maxwell's equations:

Proof. This derivation is standard but will be useful in what follows. Let $F \equiv B + dt \wedge E$, and define $\mathcal{A} \equiv (\mathbf{A}, -\phi)$, where ϕ is the $\mathbb{R}$-valued electric potential and $\mathbf{A}$ the $\mathbb{R}^3$-valued magnetic potential. Define the components of the magnetic field by $B = B_1 dx^2 \wedge dx^3 + B_3 dx^1 \wedge dx^2 + B_2 dx^3 \wedge dx^1$, and using $E = E_i dx^i$, we find $dt \wedge E = -E_i dx^i \wedge dt$. With $t \equiv x^4$, as a matrix of components we have

$$F_{ij} = \begin{pmatrix} 0 & B_3 & -B_2 & -E_1 \\ -B_3 & 0 & B_1 & -E_2 \\ B_2 & -B_1 & 0 & -E_3 \\ E_1 & E_2 & E_3 & 0 \end{pmatrix}.$$

grangian) is the action

$$S(q_0, q, t) = \int_0^t \mathcal{L}(\dot{\gamma}(\tau))\mathrm{d}\tau$$

where $\gamma : [0, t] \to T\mathcal{M}$ with $\gamma(0) = q_0, \gamma(t) = q_1$ extremises the action.

Now $d\mathcal{A} = \partial_j \mathcal{A}_i dx^j \wedge dx^i = \partial_j \mathcal{A}_i \big(dx^j \otimes dx^i - dx^i \otimes dx^j\big) = \sum_{i \neq j} \big(\partial_j \mathcal{A}_i - \partial_i \mathcal{A}_j\big) dx^i \otimes dx^j$, thus $(d\mathcal{A})_{ij} = \partial_j \mathcal{A}_i - \partial_i \mathcal{A}_j$. Hence $F = d\mathcal{A}$ implies

$$E_i = -F_{i4} = -\partial_t \mathbf{A_i} - \partial_{\mathbf{i}} \phi \qquad \text{for } \mathbf{i} = \mathbf{1}, \mathbf{2}, \mathbf{3}$$

and

$$B_i = -\tfrac{1}{2} \epsilon_{ijk} F_{jk} = -\tfrac{1}{2} \epsilon_{ijk} \big(\partial_k \mathbf{A_j} - \partial_{\mathbf{j}} \mathbf{A_k}\big) = -\epsilon_{\mathbf{ijk}} \partial_{\mathbf{k}} \mathbf{A_j} = (\nabla \times \mathbf{A})_{\mathbf{i}}$$

we recover the standard relation between $\mathbf{E} \equiv (E_1, E_2, E_3), \mathbf{B} \equiv (B_1, B_2, B_3)$ and the vector potential

$$\mathbf{E} = -\nabla \phi - \frac{\partial \mathbf{A}}{\partial \mathbf{t}} \qquad \mathbf{B} = \nabla \times \mathbf{A}.$$

Moreover the equation $dF = dd\mathcal{A} = 0$ simply yields Maxwell's equations

$$\nabla \cdot \mathbf{B} = \nabla \cdot \nabla \times \mathbf{A} = \mathbf{0} \qquad \nabla \times \mathbf{E} = -\nabla \times \nabla \phi - \nabla \times \frac{\partial \mathbf{A}}{\partial \mathbf{t}} = -\frac{\partial \mathbf{B}}{\partial \mathbf{t}}.$$

■

Derivation of Hamilton's equation for $\omega_A \equiv \omega - \mu \mathrm{d}A$ when a field $F = \mathrm{d}A$ is present (note if $\omega_A \equiv \omega - \mu F$, then $\mathrm{d}F = 0$, and thus the field field F must be exact on contractible manifolds, $F = \mathrm{d}A$):

Proof. Hamilton's equation reads $\omega_{\mathcal{A}}(X, Y) = dH(Y)$, for any $Y \in \Gamma(TT^*\mathcal{M})$ where X is the Hamiltonian vector field of H. Consider local cotangent-lifted coordinates (x^i, p_i) of $T^*\mathcal{M}$. Let $X = X^i \partial_{x^i} + \bar{X}^i \partial_{p_i}$, $Y = Y^i \partial_{x^i} + \bar{Y}^i \partial_{p_i}$. Then $dH(Y) = \partial_{x^i} H dx^i(Y) + \partial_{p_i} H dp_i(Y) = \partial_{x^i} H Y^i + \partial_{p_i} H \bar{Y}^i$. Using $dx^i \wedge dp_i = dx^i \otimes dp_i - dp_i \otimes dx^i$, we find $dx^i \wedge dp_i(X, Y) = X^i \bar{Y}^i - \bar{X}^i Y^i$. Moreover

$$\begin{aligned} \tau^* d\mathcal{A}(X, Y) &= \partial_{x^k} \mathcal{A}_i dx^k \wedge dx^i(\partial\tau(X), \partial\tau(Y)) = \partial_{x^k} \mathcal{A}_i dx^k \wedge dx^i(X^r \partial_{x^r}, Y^s \partial_{x^s}) \\ &= Y^s X^r \partial_{x^k} \mathcal{A}_i dx^k \wedge dx^i(\partial_{x^r}, \partial_{x^s}) = Y^s X^r \big(\partial_s \mathcal{A}_r - \partial_r \mathcal{A}_s\big). \end{aligned}$$

Hence, comparing coefficients of Y gives

$$\bar{X}^j = -\partial_{x^j} H - e X^i \big(\partial_i \mathcal{A}_j - \partial_j \mathcal{A}_i\big) \qquad X^i = \partial_{p_i} H.$$

Now consider $\mathcal{M} = \mathbb{R}^{3+1}$ as in Maxwell's equations above. Let $i, j \in \{1, 2, 3\}, \mu, \nu \in \{1, 2, 3, 4\}$ Then

$$X^\mu \big(\partial_\mu \mathcal{A}_j - \partial_j \mathcal{A}_\mu\big) = X^i \partial_i \mathcal{A}_j + X^t \partial_t \mathcal{A}_j - X^i \partial_j \mathcal{A}_i - X^t \partial_j \mathcal{A}_t$$

A solution $\gamma : \mathbb{R} \to \mathbb{R}^3$ of Hamilton's equations satisfy $\dot{\gamma}^i = X^i$, and we have $X^t = \frac{dt}{dt} = 1$.

Then

$$X^t\partial_t\mathcal{A}_i - X^t\partial_{x^i}\mathcal{A}_t = \partial_t\mathcal{A}_i - \partial_{x^i}\mathcal{A}_t$$

which, from the derivation of Maxwell's equations above, are the components of the vector $\partial_t\mathbf{A} + \nabla\phi = -\mathbf{E}$. Similarly $X^i\partial_i\mathcal{A}_j - X^i\partial_j\mathcal{A}_i$ are the components of the vector $-\dot{\gamma}\times\mathbf{B}$ since

$$(\dot{\gamma}\times\mathbf{B})_i = \epsilon_{ijk}\dot{\gamma}_j B_k = \epsilon_{ijk}\epsilon_{krs}\dot{\gamma}_j\partial_r A_s = (\delta_{ir}\delta_{js} - \delta_{is}\delta_{jr})\dot{\gamma}_j\partial_r A_s = \dot{\gamma}_s\partial_i A_s - \dot{\gamma}_r\partial_r A_i.$$

Thus Hamilton's equations read

$$\frac{dp}{dt} = -d_x H + e\big(\dot{\gamma}\times\mathbf{B} + \mathbf{E}\big)$$

■

A.5.7.2 Representations

Suppose now that $\mathcal{G}$ is not a matrix group, and let (e_i) be a basis of left-invariant vector fields associated to a basis (ξ_i) of $\mathfrak{g}$. Fix an inner product on $\mathfrak{g}$ and define $G_{ij} \equiv \langle\xi_i,\xi_j\rangle_{\mathfrak{g}}$. The Hamiltonian vector fields corresponding to T and V on $\mathcal{G}\times\mathfrak{p}$ are $\hat{T}(g,X) = \partial_1 L_g X$ and $\hat{V}(g,X) = -G^{jk}e_j|_g(V)\xi_k$ (see §3.9.1.2). In order to obtain matrix-valued Hamiltonian trajectories on a submanifold of $\mathbb{R}^{n\times n}$ we use a faithful (injective) representation $\psi : \mathcal{G}\to \mathrm{GL}(n,\mathbb{R})$ of the Lie group whose tangent map $\partial_1\psi : \mathfrak{g}\to\mathbb{R}^{n\times n}$ is injective.

The potential energy V defines a function $V\circ\psi^{-1}$ on the image of the representation $\psi(\mathcal{G})$. If $\psi(\mathcal{G})$ is not an open subset of $\mathrm{GL}(n)$, it will be useful to use an arbitrary smooth extension $V_\psi : \mathrm{GL}(n)\to\mathbb{R}$ of $V\circ\psi^{-1}$ (so $V_\psi = V\circ\psi^{-1}$ on $\psi(\mathcal{G})$) in order to differentiate the potential energy with respect to global coordinates. Using ψ we also construct the matrix generators $T_i\in\mathbb{R}^{n\times n}$ associated with the Lie algebra basis elements $\xi_i\in\mathfrak{g}$, by setting $T_i \equiv \partial_1\psi(\xi_i)$. We then denote by $P\equiv v^iT_i$ the matrix element representing the momentum $v^i\xi_i$. Its kinetic energy $T_\psi : \partial_1\psi(\mathfrak{g})\subseteq\mathbb{R}^{n\times n}\to\mathbb{R}$ is given by $T_\psi(P)\equiv\frac{1}{2}G_{ij}v^iv^j$. By construction we have $T = T_\psi\circ\partial_1\psi$ and $V = V_\psi\circ\psi$. We summarise these definitions in Table A.1.

The Hamiltonian trajectories on $\mathcal{G}\times\mathfrak{g}$ are the integral curves $t\mapsto(q(t),p(t))\in\mathcal{G}\times\mathfrak{g}$ of the vector field $\hat{H} = \hat{T}+\hat{V}$ which are solutions to Hamilton's equations

$$(\dot{q}(t),\dot{p}(t)) = \hat{H}(q,p) = \hat{T}(q,p) + \hat{V}(q,p) = (\partial_1 L_q p, 0) + \big(0, -G^{jk}e_j(V)|_q\xi_k\big)$$

i.e.,

$$\dot{q}(t) = \partial_1 L_q p, \qquad \dot{p}(t) = -G^{jk}e_j(V)|_q\xi_k. \tag{A.12}$$

Hamiltonian System	$\mathcal{G} \times \mathfrak{g}$	$\mathrm{GL}(n) \times \mathbb{R}^{n\times n}$
Position	$q \in \mathcal{G}$	$Q \in \psi(\mathcal{G}) \subset \mathrm{GL}(n)$
Generators	basis $\xi_i \in \mathfrak{g}$	$T_i \equiv \partial_1\psi(\xi_i) \in \mathbb{R}^{n\times n}$
Momentum/velocity	$p = v^i\xi_i \in \mathfrak{g}$	$P = v^iT_i \in \mathbb{R}^{n\times n}$
Potential Energy	$V : \mathcal{G} \to \mathbb{R}$	$V_\psi : \mathrm{GL}(n) \to \mathbb{R}$
Kinetic Energy	$T(p) = \frac{1}{2}G_{ij}v^iv^j$	$T_\psi(P) = \frac{1}{2}G_{ij}v^iv^j$

Table A.1: Hamiltonian system on $\mathcal{G} \times \mathfrak{g}$ and its image under the representation ψ. The equations of motion on $\mathcal{G} \times \mathfrak{g}$ are given by (A.12), and their image on $\mathrm{GL}(n) \times \mathbb{R}^{n\times n}$ satisfies (A.13).

The associated equations in terms of matrix trajectories $(Q, P) \equiv (\psi \circ q, \partial_1\psi \circ p)$ on $\mathrm{GL}(n) \times \mathbb{R}^{n\times n}$ are (see §3.9.1.3)

$$\dot{Q}(t) = Q(t) \cdot P(t), \qquad \dot{P}(t) = -\operatorname{tr}\big(\partial_x V_\psi \cdot Q(t) \cdot T_i\big)T^i, \tag{A.13}$$

where $\cdot$ denotes the matrix product, and $T^i \equiv G^{ij}T_j$. The rate of change of momentum P and position Q along an integral curve of $\hat{V}$ is

$$Q(t) = Q(0), \qquad \dot{P}(t) = -\operatorname{tr}\big(\partial_x V_\psi \cdot Q(0) \cdot T_i\big)T^i;$$

whose solution

$$Q(t) = Q(0), \qquad P(t) = P(0) - t\operatorname{tr}\big(\partial_x V_\psi \cdot Q(0) \cdot T_i\big)T^i.$$

Along an integral curve of $\hat{T}$ we have

$$P(t) = P(0), \qquad \dot{Q}(t) = Q(t) \cdot P(0).$$

which implies

$$Q(t) = Q(0) \cdot \exp\big(tP(0)\big), \qquad P(t) = P(0).$$

A.5.7.3 Symplectomorphism $T^*\mathcal{G} \cong \mathcal{G} \times \mathfrak{g}^*$

The cotangent bundle over $\mathcal{G}$ is isomorphic to a trivial bundle $T^*\mathcal{G} \cong \mathcal{G} \times \mathfrak{g}^* \cong \mathcal{G} \times \mathbb{R}^n$. Indeed fix a frame $e_i \in \mathfrak{g}_L$ and co-frame $\theta^i \in \mathfrak{g}_L^*$. We have the maps

$$(g, p) = (g, p_i\theta^i_g) \mapsto (g, p_i\theta^i_1) \mapsto \big(g, (p_1, \ldots, p_n)\big)$$

Letting $F : T^*\mathcal{G} \to \mathcal{G} \times \mathfrak{g}^*$ be this diffeomorphism. Then

$$\big(F^*\tilde{p}_i\big)(g,\alpha_g) = (\tilde{p}_i \circ F)(g,\alpha_g) = \tilde{p}_i \circ (F(g,\alpha_g)) = \tilde{p}_i \circ (g, L_g^*\alpha_g) = L_g^*\alpha_g(e_i(1)) = p_i(\alpha_g)$$

and

$$F^*(\pi^*\theta^i) = \theta^i \circ \partial\pi \circ \partial F = \theta^i \circ \partial(\pi \circ F) = \theta^i \circ \partial\pi = \pi^*\theta^i.$$

Hence $T^*\mathcal{G}$ with canonical symplectic structure is symplectomorphic to $\mathcal{G} \times \mathfrak{g}^*$ with the symplectic structure from [187, 186].

A.5.7.4 Symplectic Structure on $T\mathcal{G}$

The pull-back of the coordinate functions p_i is $p_i \circ \flat : T\mathcal{G} \to \mathbb{R}$, and

$$\begin{aligned} p_i \circ \flat(h, v_h) &= p_i\big(\langle v_h, \cdot\rangle\big) = \langle v_h, e_i(h)\rangle = v_h^j \langle e_j(h), e_i(h)\rangle \\ &= v_h^j \langle e_j, e_i\rangle(h) = \langle e_j, e_i\rangle(h)\theta_h^j(v_h) = \big(\theta^j \pi_{T\mathcal{G}}^* \langle e_j, e_i\rangle\big)(h, v_h), \end{aligned}$$

where $\pi_{T\mathcal{G}} : T\mathcal{G} \to \mathcal{G}$ is the canonical projection. Note in the final expression θ^j is viewed as a map $T\mathcal{G} \to \mathbb{R}$ (which we denote by v^j) rather than an element of $\Gamma(T^*\mathcal{G})$. This gives the definition of the coordinate functions v^j on $T\mathcal{G}$. On the other hand, $\flat^*\pi^*\theta^i = (\pi \circ \flat)^*\theta^i$ and $\pi \circ \flat : T\mathcal{G} \to \mathcal{G}$ is simply $\pi_{T\mathcal{G}}$. In summary,

$$\flat^*\big(p_i\pi^*\theta^i\big) = \big(p_i \circ \flat\big)\flat^*\pi^*\theta^i = \big(v^j \pi_{T\mathcal{G}}^* \langle e_j, e_i\rangle\big)\pi_{T\mathcal{G}}^*\theta^i = v^j \pi_{T\mathcal{G}}^*\big(\langle e_j, e_i\rangle\, \theta^i\big)$$

A.5.7.5 Symplectomorphism Left-Trivialisation

The isomorphism $T\mathcal{G} \cong \mathcal{G} \times \mathfrak{g}$ can be written as $(g, v) \mapsto (g, \theta_g(v)) = (g, \partial_g L_{g^{-1}} v)$, where θ is the (left) Maurer-Cartan form. We can thus pull-back our Hamiltonian system from $T\mathcal{G}$ to $\mathcal{G} \times \mathfrak{g}$: let $f : \mathcal{G} \times \mathfrak{g} \to T\mathcal{G}$ be this map, and let $E = V \circ \pi + \frac{1}{2}\langle \cdot, \cdot \rangle$ be an energy function on $T\mathcal{G}$. Then

$$f^*E = E \circ f = V \circ \pi \circ f + \tfrac{1}{2}\langle \cdot, \cdot \rangle \circ f = V \circ \pi + \tfrac{1}{2}\langle f(\cdot), f(\cdot)\rangle$$

and

$$\begin{aligned} \big\langle f(g, v^i e_1(1)), f(g, v^j e_j(1))\big\rangle &= \big\langle (g, v^i e_i(g)), (g, v^j e_j(g))\big\rangle = \big\langle v^i e_i(g), u^j e_j(g)\big\rangle_g \\ &= L_g^* \big\langle v^i e_i(1), u^j e_j(1)\big\rangle \end{aligned}$$

Thus if the metric is left-invariant

$$f^*E = V \circ \pi + \tfrac{1}{2}\langle \cdot, \cdot \rangle_1 .$$

A.5.7.6 Hamiltonian Vector Fields

We proceed similarly to [187] but work on $\mathcal{G} \times \mathfrak{g}$ which will enable us to identify the coordinate vector fields ∂_{v^i} with ξ_i. An arbitrary vector field Y on $\mathcal{G} \times \mathfrak{g}$ may be expanded as $Y = Y^i e_i + \bar{Y}^i \partial_{v^i}$, where $v^i : (g, u_1) \mapsto \theta_1^i(u_1)$ are the fibre coordinates. This follows because any vector tangent to $\mathcal{G} \times \mathfrak{g}$ is given by the derivative of a curve $\gamma : [a, b] \to \mathcal{G} \times \mathfrak{g}$ which may be written as $\gamma(t) = (q(t), u(t))$; then $\dot{\gamma} = (\dot{q}, \dot{u})$, where $\dot{q} \in T\mathcal{G}$ and $\dot{u} \in T\mathfrak{g}$ can be expanded in terms of the (global trivialising) frames e_i and ∂_{v^i}.

The equation $\omega_\flat(\hat{f}, Y) = \mathrm{d}f(Y)$ must hold for every Y if $\hat{f}$ is the Hamiltonian vector field of f. Note

$$\pi^*\theta^i \wedge \mathrm{d}v^j = \pi^*\theta^i \otimes \mathrm{d}v^j - \mathrm{d}v^j \otimes \pi^*\theta^i$$

and set $\hat{f} = X^i e_i + \bar{X}^i \partial_{v^i}$ Then $\pi^*\theta^i \wedge \mathrm{d}v^j(\hat{f}, Y) = X^i \bar{Y}^j - Y^i \bar{X}^j$, moreover $\pi^*\theta^i \wedge \pi^*\theta^k(\hat{f}, Y) = X^j Y^k - Y^j X^k$, and finally $\mathrm{d}f(Y) = Y(f) = Y^i e_i(f) + \bar{Y}^i \partial_{v^i} f$. Hence $\omega_{\mathcal{L}}(\hat{f}, Y) = \mathrm{d}f(Y)$ is equivalent to

$$G_{ij}\big(X^i \bar{Y}^j - Y^i \bar{X}^j\big) + \tfrac{1}{2} v^r c_{rjk}\big(X^j Y^k - Y^j X^k\big) = Y^i e_i(f) + \bar{Y}^i \partial_{v^i} f$$

where $c_{rjk} \equiv G_{rl} c^l_{jk}$. As this must hold for every Y, equating the coefficients of Y^i and $\bar{Y}^i$ yields $G_{ij} X^i = \partial_{v^j} f$ and $G_{ij} \bar{X}^j = v^r c_{rji} X^j - e_i(f)$ respectively, where we have used $c_{rjk} = -c_{rkj}$ since the Lie bracket of vector field is antisymmetric: $[Z, W] = -[W, Z]$.

It follows that

$$\hat{f} = G^{jk} \partial_{v^k} f e_j + \big(G^{ik} G^{lj} v^r c_{rji} \partial_{v^l} f - G^{jk} e_j(f)\big) \partial_{v^k}.$$

In particular if $T = \frac{1}{2} G_{ij} v^i v^j$ and V is independent of fibre coordinates v^i, then

$$\hat{T} = v^k e_k + G^{ik} v^r v^s c_{rsi} \partial_{v^k} \qquad \hat{V} = -G^{jk} e_j(V) \partial_{v^k}$$

where we have used $e_i[T] = 0$ since $e_i[T] = \frac{1}{2} e_i\big(G_{jk}(g)\big) v^j v^k = \frac{1}{2} e_i\big(G_{jk}(1)\big) v^j v^k = 0$ by left-invariance of the metric.

In fact the coordinate vectors ∂_{v^k} are simply the generators, $\partial_{v^k} = \xi_k$, since both of them have the integral curve $\gamma : t \mapsto t\xi_k$. Indeed, if $v = (v^1, \ldots, v^n) : \mathfrak{g} \to \mathbb{R}^n$ denotes the coordinate chart $\dot{\gamma}[f] = \frac{\partial (f \circ v^{-1})}{\partial v^i} \frac{\partial v^i \circ \gamma}{dt}|_{t=0} = \frac{\partial (f \circ v^{-1})}{\partial v^k} = \frac{\partial}{\partial v^k}(f)$.

Note that for a general f, $e_j(f)$ is a function on $T\mathcal{G}$, however since V does not depend on

the fibre coordinates, $e_j(V)$ is just a function of $\mathcal{G}$.

A.6 Additional Material Diffusions

A.6.1 Overdamped Langevin Processes

Previously we have focused on the geometry of underdamped Langevin processes on phase space $\mathcal{F} = T\mathcal{M}$. We shall now discuss overdamped Langevin processes on configuration space $\mathcal{F} = \mathcal{M}$. Consider a target $P = e^{-\beta V}\mu_{\mathcal{M}}$. As mentioned earlier Langevin processes are characterised by the fact they are measure-preserving yet contain no rotationnel terms: both $\text{curl}_{\mu_{\mathcal{M}}(\mathcal{A})}$ and $\text{div}_{\mu_{\mathcal{M}}}(Y_i)Y_i$ in (4.28) do not appear. Underdamped Langevin systems further required the noise to be entirely vertical. We may think of the overdamped case as instead further requiring that the generator is L^2-symmetric, that is

$$\mathrm{d}Q_t = -\frac{\beta}{2}Y_i(V)Y_i\mathrm{d}t + Y_i \circ \mathrm{d}W_t^t,$$

where the pair $(\{Y_i\}_i, \mu_{\mathcal{M}})$ satisfies $\text{div}_{\mu_{\mathcal{M}}}(Y_i)Y_i = 0$, which ensures the above is P-preserving.

To relate this to the standard overdamped Langevin equation, consider a Riemannian manifold $\mathcal{M}$. The Riemannian Brownian motion can be obtained as a projection of a diffusion in Stratonovich form. Specifically, if we define the canonical horizontal vector field L_i on the orthonormal frame bundle $\pi_O : O(\mathcal{M}) \to \mathcal{M}$, then the diffusion $\mathrm{d}O_t = L_i \circ \mathrm{d}W_t^i$ reduces to the Riemannian Brownian motion (see theorem 4.2 [176]). Moreover $L_i(\pi_O^* V)L_i$ is π_O-related to $\nabla^{\mathcal{M}}V$, as follows from equations (4.12) and (4.22) [176]. [16] Hence the usual Riemannian overdamped Langevin process which is used in MALA (§4.1.2)

$$\mathrm{d}Q_t = -\frac{\beta}{2}\nabla^{\mathcal{M}}V\mathrm{d}t + \mathrm{d}B_t,$$

where B_t is the Riemannian Brownian (generated by the Laplace–Beltrami operator), is the projection under π_O of the Stratonovich diffusions $\mathrm{d}Q_t = -\frac{\beta}{2}L_i(\pi_O^* V)L_i\mathrm{d}t + L_i \circ \mathrm{d}W_t^t$,

Finally we mention that if X_t is a Riemannian brownian motion and T is a Riemannian submersion, then $T(X_t)$ is a Riemannian brownian motion. In particular for the naturally reductive homogeneous manifolds we considered in the previous chapter, $\tau : \mathcal{G} \to \mathcal{G}/\mathcal{K} = \mathcal{M}$ is

[16] Locally $L_i = e_i^r\partial_{q^r} - \Gamma_{sa}^{\ell}e_i^s e_b^a \partial_{e_b^\ell}$ where $(e_s)_s$ is an orthonormal frame, so $L_i(\pi_O^* V)L_i = e_i^r\partial_{q^r}(V)e_i^s\partial_{q^s} = G^{rs}\partial_{q^r}(V)\partial_{q^s}$.

a Riemannian submersion, hence

$$\mathrm{d}g_t = -\frac{\beta}{2}\nabla^{\mathcal{G}}\tau^* V \mathrm{d}t + \mathrm{d}B_t^{\mathcal{G}}$$

reduces to the Riemannian overdamped Langevin process on $\mathcal{M}$, so that as we did in HMC, we can implement the process on the Lie group.

A.7 Additional Material on DKSD

A.7.1 $\mathcal{B}$-KSD

Recall the notation of §6.9 and fix a bracket $\mathcal{B}$. We briefly discuss the extension of KSD to manifolds. Let $\Phi_k : \mathcal{F} \to \mathcal{H}_k$ be the meaning embedding of a smooth scalar kernel k on $\mathcal{F}$. We say that the kernel k is in the $\mathcal{B}$-Stein class of μ if the μ-Pettis-integral in $\mathcal{H}_k$ of $\mathrm{div}_\mu(Y_{\Phi_k}) : \mathcal{F} \to \mathcal{H}_k$ vanishes (the latter function is defined as $x \mapsto \mathrm{div}_\mu(Y_{k_x})$).

Theorem 64. *Suppose that Casimir$(\mathcal{B}) \bigcap \mathcal{H}_k = \{0\}$, that k is in the $\mathcal{B}$-Stein class of μ, and that $div_\mu(Y_h) = \langle h, div_\mu(Y_{\Phi_k})\rangle_k$ for all $h \in \mathcal{H}_k$. In the setting of theorem 49, the direction of greatest change of KL within $\mathfrak{X}_{\mathcal{B}}|_{\mathcal{H}_k}$*

$$\sup_{Y \in \mathfrak{X}_{\mathcal{B}}|_{\mathcal{H}_k}, \|Y\|=1} \left(-\frac{\mathrm{d}}{\mathrm{d}t} KL(Q\|\mu)\right) = \sup_{f \in \mathcal{H}_k, \|f\|_k = 1} \int \mathcal{T}_\mu^{\mathcal{B}}(f) Q$$

is achieved at $\int div_\mu(Y_{\Phi_k})Q \in \mathcal{H}_k$, provided the function $div_\mu(Y_{\Phi_k})$ is Q-Pettis-integrable in $\mathcal{H}_k$. We define the ***bracket kernelised Stein discrepancy*** *as the square of the RHS:*

$$\mathcal{B}\text{-}KSD(Q\|\mu) \equiv \sup_{f \in \mathcal{H}_k, \|f\|_k = 1} \left| \int \mathcal{T}_\mu^{\mathcal{B}}(f) Q \right|^2 = \int\int k_\mu^{\mathcal{B}}(x,y) Q(x) Q(y), \tag{A.14}$$

where we have defined the ***$\mathcal{B}$-Stein kernel***

$$k_\mu^{\mathcal{B}}(x,y) \equiv \mathcal{T}_\mu^{\mathcal{B}}(\mathcal{T}_\mu^{\mathcal{B}}(k_x))(y) = div_\mu(X_{div_\mu(X_{k_x})})(y). \tag{A.15}$$

Proof. By theorem 49 and proposition 50 we have

$$\sup_{Y \in \mathfrak{X}_{\mathcal{B}}|_{\mathcal{H}_k}, \|Y\|=1} \int \mathrm{div}_\mu(Y) Q = \sup_{f \in \mathcal{H}_k, \|f\|_k = 1} \int \mathrm{div}_\mu(Y_f) Q.$$

Using the assumption $\mathrm{div}_\mu(Y_h) = \langle h, \mathrm{div}_\mu(Y_{\Phi_k})\rangle_k$, we further have

$$\sup_{h\in\mathcal{H}_k,\|h\|_k=1}\int \mathrm{div}_\mu(Y_h)Q = \sup_{h\in\mathcal{H}_k,\|h\|_k=1}\int \langle h, \mathrm{div}_\mu(Y_{\Phi_k})\rangle_k\, Q$$
$$= \sup_{h\in\mathcal{H}_k,\|h\|_k=1}\left\langle h, \int \mathrm{div}_\mu(Y_{\Phi_k})Q\right\rangle_k = \|\int \mathrm{div}_\mu(Y_{\Phi_k})Q\|_k.$$

In particular since k is in the $\mathcal{B}$-Stein class of μ, then $\mathcal{H}_k \subset C_\mu$, since

$$\int \mathrm{div}_\mu(Y_h)\mu = \left\langle h, \int \mathrm{div}_\mu(Y_{\Phi_k})\mu\right\rangle_k = 0,$$

so $\mathrm{div}_\mu \circ \mathcal{B}^\sharp \circ \mathrm{d} = \mathcal{T}_\mu^{\mathcal{B}}$ on $\mathcal{H}_k$. Finally

$$\mathcal{B}\text{-KSD}(Q\|\mu) \equiv \sup_{f\in\mathcal{H}_k,\|f\|_k=1}\left|\int \mathrm{div}_\mu(Y_f)Q\right|^2 = \left\langle \int \mathrm{div}_\mu(Y_{\Phi_k})Q, \int \mathrm{div}_\mu(Y_{\Phi_k})Q\right\rangle_k$$
$$= \int\int \langle \mathrm{div}_\mu(Y_{\Phi_k})(x), \mathrm{div}_\mu(Y_{\Phi_k})(y)\rangle_k\, Q(x)Q(y)$$
$$= \int\int \mathrm{div}_\mu(Y_{\mathrm{div}_\mu(Y_{\Phi_k})(x)})(y)Q(x)Q(y) = \int\int \mathrm{div}_\mu(Y_{\mathrm{div}_\mu Y_{k(x,\cdot)}})(y)Q(x)Q(y).$$

■

We note that the assumption $\mathrm{div}_\mu(Y_h) = \langle h, \mathrm{div}_\mu(Y_{\Phi_k})\rangle_k$ for all $h \in \mathcal{H}_k$ holds for example when $\mathcal{T}_\mu^{\mathcal{B}}$ is the diffusion Stein operator (6.19) of a diffusion whose semigroup is strongly continuous by proposition 1 [154]. The latter holds if the diffusion satisfies Hörmander condition, since the associated semigroup is strongly Feller [18].

A.7.2 The Stein Kernel

Note the Stein kernel satisfies

$$k^0 = \frac{1}{p(y)p(x)}\nabla_y \cdot \nabla_x \cdot \left(p(x)m(x)Km(y)^\top p(y)\right)$$

since

$$k^0 = \mathcal{S}_p^2\mathcal{S}_p^1 K(x,y) = \frac{1}{p(y)p(x)}\nabla_y \cdot (p(y)m(y)\nabla_x \cdot (p(x)m(x)K))$$
$$= \frac{1}{p(y)p(x)}\nabla_y \cdot (p(y)m(y)\partial_{x^i}\left(p(x)m(x)_{ir}K_{rs}\right)e_s)$$
$$= \frac{1}{p(y)p(x)}\nabla_y \cdot (p(y)m(y)_{ls}\partial_{x^i}\left(p(x)m(x)_{ir}K_{rs}\right)e_l)$$

$$
\begin{aligned}
&= \frac{1}{p(y)p(x)} \partial_{y^l} \left(p(y) m(y)_{ls} \partial_{x^i} \left(p(x) m(x)_{ir} K_{rs} \right) \right) \\
&= \frac{1}{p(y)p(x)} \partial_{y^l} \partial_{x^i} \left(p(x) m(x)_{ir} K_{rs} m(y)^\top_{sl} p(y) \right) \\
&= \frac{1}{p(y)p(x)} \nabla_y \cdot \nabla_x \cdot \left(p(x) m(x) K m(y)^\top p(y) \right).
\end{aligned}
$$

We can expand the Stein kernel using the following expressions:

$$
\begin{aligned}
&\nabla_y \cdot (p(y) m(y) \nabla_x \cdot (p(x) m(x) K)) \\
&= \nabla_y \cdot \left(p(y) m(y) \left(K m(x)^\top \nabla_x p + p(x) \nabla_x \cdot (m(x) K) \right) \right).
\end{aligned}
$$

$$
\begin{aligned}
&\nabla_y \cdot \left(p(y) m(y) K m(x)^\top \nabla_x p \right) \\
&= m^\top(x) \nabla_x p \cdot K m(y)^\top \nabla_y p + p(y) \nabla_y \cdot \left(m(y) K m(x)^\top \nabla_x p \right) \\
&= m^\top(x) \nabla_x p \cdot K m(y)^\top \nabla_y p + p(y) \nabla_y \cdot (m(y) K) \cdot m(x)^\top \nabla_x p,
\end{aligned}
$$

$$
\begin{aligned}
&\nabla_y \cdot (p(y) m(y) p(x) \nabla_x \cdot (m(x) K)) \\
&= p(x) \left(\nabla_y \cdot (p(y) m(y)) \cdot \nabla_x \cdot (m(x) K) + p(y) \operatorname{tr} \left[m(y) \nabla_y \nabla_x \cdot (m(x) K) \right] \right) \\
&= p(x) p(y) \operatorname{tr} \left[m(y) \nabla_y \nabla_x \cdot (m(x) K) \right] \\
&\quad + p(x) \nabla_x \cdot (m(x) K) \cdot \left(m(y)^\top \nabla_y p + p(y) \nabla_y \cdot m \right).
\end{aligned}
$$

Hence

$$
\begin{aligned}
k^0 &= m^\top(x) \nabla_x \log p \cdot K m(y)^\top \nabla_y \log p \\
&\quad + \nabla_y \cdot (m(y) K) \cdot m(x)^\top \nabla_x \log p + \nabla_x \cdot (m(x) K) \cdot m(y)^\top \nabla_y \log p \\
&\quad + \nabla_x \cdot (m(x) K) \cdot \nabla_y \cdot m + \operatorname{tr} \left[m(y) \nabla_y \nabla_x \cdot (m(x) K) \right] \\
&= \langle s_p(x), K s_p(y) \rangle + \langle \nabla_y \cdot (m(y) K), s_p(x) \rangle + \langle \nabla_x \cdot (m(x) K), s_p(y) \rangle \\
&\quad + \langle \nabla_x \cdot (m(x) K), \nabla_y \cdot m \rangle + \operatorname{tr} \left[m(y) \nabla_y \nabla_x \cdot (m(x) K) \right]
\end{aligned}
$$

A.7.3 Special Cases of Diffusion Kernel Stein Discrepancy

Consider

$$
k^0 = \frac{1}{p(y)p(x)} \nabla_y \cdot \nabla_x \cdot \left(p(x) m(x) K(x, y) m(y)^\top p(y) \right)
$$

and decompose $m(x)K(x,y)m(y)^\top \equiv gA$ where g is scalar and A is matrix-valued. Then we

$$\begin{aligned}k^0 &= g\left\langle \nabla_y \log p, A\nabla_x \log p\right\rangle + \left\langle \nabla_y \log p, A\nabla_x g\right\rangle + \left\langle \nabla_y g, A\nabla_x \log p\right\rangle \\ &+ \mathrm{Tr}[A\nabla_x\nabla_y g] + g\nabla_y\cdot\nabla_x\cdot A + \left\langle \nabla_x\cdot A, \nabla_y g\right\rangle + \left\langle \nabla_y\cdot A^\top, \nabla_x g\right\rangle \\ &+ g\left\langle \nabla_y\cdot A^\top, \nabla_x \log p\right\rangle + g\left\langle \nabla_x\cdot A, \nabla_y \log p\right\rangle.\end{aligned}$$

For the case, $K = \mathrm{diag}(k^1,\ldots,k^d)$, setting $\mathcal{T}_i^x \equiv \frac{1}{p(x)}\partial_{x^i}\left(p(x)\cdot\right)$ then

$$\mathcal{S}_p^2\mathcal{S}_p^1[\mathrm{diag}(k^1,\ldots,k^d)] = \mathcal{T}_l^y\left(m_{li}(y)\mathcal{T}_c^x\left(k^i(x,y)m_{ic}^\top(x)\right)\right) = \mathcal{T}_l^y\mathcal{T}_c^x\left(m_{li}(y)k^i(x,y)m_{ci}(x)\right).$$

If $K = Ik$ in components

$$\begin{aligned}\mathcal{S}_p^2\mathcal{S}_p^1[Ik] &= (s_p(x))_i k(x,y)(s_p(y))_i + \partial_{y^i}(m_{ir}k)(s_p(x))_r + \partial_{x^i}(m(x)_{ir}k)(s_p(y))_r \\ &+ \partial_{x^i}(m(x)_{ir}k)\partial_{y^l}(m_{lr}) + m(y)_{ir}\partial_{y^i}\partial_{x^s}(m(x)_{sr}k)\end{aligned}$$

When $p = p_\theta$ we are often interested in the gradient $\nabla_\theta k_\theta^0$. Note $\nabla_y\cdot(m(y)K) = k\nabla_y\cdot m + \nabla_y k\cdot m(y)$, so [17]

$$\begin{aligned}\partial_{\theta^i}\left[k\left\langle \nabla_y\cdot m, s_p(x)\right\rangle\right] &= k\partial_{\theta^i}\left\langle \nabla_y\cdot m, s_p(x)\right\rangle \\ \partial_{\theta^i}\left[\left\langle \nabla_y k\cdot m(y), s_p(x)\right\rangle\right] &= \left\langle \nabla_y k, \partial_{\theta^i}[m(y)s_p(x)]\right\rangle \\ \mathrm{tr}\left[m(y)\nabla_y\nabla_x\cdot(m(x)K)\right] &= \nabla_y k^\top m(y)\nabla_x\cdot m + \mathrm{Tr}[m(y)m(x)^\top\nabla_y\nabla_x k]\end{aligned}$$

and the terms in $\partial_{\theta^i}k^0$ reduce to

$$\begin{aligned}\partial_{\theta^i}\left\langle s_p(x), Ks_p(y)\right\rangle &= k\partial_{\theta^i}\left\langle s_p(x), s_p(y)\right\rangle \\ \partial_{\theta^i}\left\langle \nabla_y\cdot(m(y)K), s_p(x)\right\rangle &= k\partial_{\theta^i}\left\langle \nabla_y\cdot m, s_p(x)\right\rangle + \left\langle \nabla_y k, \partial_{\theta^i}[m(y)s_p(x)]\right\rangle \\ \partial_{\theta^i}\left\langle \nabla_x\cdot(m(x)K), s_p(y)\right\rangle &= k\partial_{\theta^i}\left\langle \nabla_x\cdot m, s_p(y)\right\rangle + \left\langle \nabla_x k, \partial_{\theta^i}[m(x)s_p(y)]\right\rangle \\ \partial_{\theta^i}\left\langle \nabla_x\cdot(m(x)K), \nabla_y\cdot m\right\rangle &= k\partial_{\theta^i}\left\langle \nabla_x\cdot m, \nabla_y\cdot m\right\rangle + \partial_{\theta^i}\left\langle \nabla_x k\cdot m(x), \nabla_y\cdot m\right\rangle.\end{aligned}$$

When $K = kI$ and we further have a diagonal matrix $m = \mathrm{diag}(f_i)$, $m(y)m(x)^\top = \mathrm{diag}(f_i(y)f_i(x))$. If $u\odot v$ denotes the vector given by the pointwise product of vectors, i.e., $(u\odot v)_i = u_i v_i$, and f is

[17] More generally $\nabla_y\cdot(m(y)K) = (\nabla_y\cdot m)\cdot K + \mathrm{Tr}[\nabla_y K\otimes m(y)]$ where $\mathrm{Tr}[\nabla_y K\otimes m]_r = \partial_{y^i}K_{jr}m_{ij}$ and if $K = Bk$

$$\begin{aligned}\partial_{\theta^i}\left[(\nabla_y\cdot m)\cdot Ks_p(x)\right] &= kB_{sr}\partial_{\theta^i}\left((\nabla_y\cdot m)_s(s_p(x))_r\right) = k\mathrm{Tr}[B\partial_{\theta^i}(s_p(x)\otimes\nabla_y\cdot m)] \\ \partial_{\theta^i}\left[\nabla_y k^\top m(y)Bs_p(x)\right] &= \partial_{y^s}kB_{jr}\partial_{\theta^i}[m_{sj}(y)(s_p(x))_r]\end{aligned}$$

the vector, then $m(x)\nabla_x \log p = f(x) \odot \nabla_x \log p$ and $(\nabla_y \cdot m)_i = \partial_{y^i} f_i$, $(\nabla_x \cdot (mk))_i = \partial_{x^i}(f_i k)$,

$$\begin{aligned} s_p(x) \cdot K s_p(y) &= k(x,y) f_i(x) \partial_{x^i} \log p f_i(y) \partial_{y^i} \log p \\ \nabla_y \cdot (m(y)K) \cdot s_p(x) &= \partial_{y^i}(f_i(y)k) f_i(x) \partial_{x^i} \log p \\ \nabla_x \cdot (m(x)K) \cdot \nabla_y \cdot m &= \partial_{x^i}(f_i(x)k) \partial_{y^i}(f_i(y)) \\ \operatorname{Tr}\left[m(y)\nabla_y \nabla_x \cdot (mk)\right] &= f_i(y)\partial_{x^i}\left(f_i(x)\partial_{y^i} k\right) \end{aligned}$$

and if $m \mapsto mI$ (is scalar), (this is just KSD with $k(x,y) \mapsto m(x)k(x,y)m(y)$):

$$\begin{aligned} k^0 = & m(x)m(y)k(x,y)\nabla_x \log p \cdot \nabla_y \log p \\ & + m(x)\nabla_y (m(y)k) \cdot \nabla_x \log p + m(y)\nabla_x (m(x)k) \cdot \nabla_y \log p \\ & + \nabla_x (m(x)k) \cdot \nabla_y m + m(y)\nabla_x \cdot (m(x)\nabla_y k), \end{aligned}$$

When $m = I$, we recover the usual definition of kernel-Stein discrepancy (KSD):

$$\operatorname{KSD}(\mathbb{Q}\|\mathbb{P})^2 = \int_{\mathcal{X}} \int_{\mathcal{X}} \frac{1}{p(y)p(x)} \nabla_y \cdot \nabla_x \left(p(x)k(x,y)p(y)\right) \mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y).$$

A.8 Minimum MMD Estimators for Generative Models

Our motivation in §5.5 for introducing the KSD was that the Stein–Langevin operator removed the integration with respect to the model distributions provided they had differentiable log-densities. However, for many statistical models of interest in modern statistical inference, the model densities are completely unknown. We then need to resort to **likelihood-free inference** methods. One popular approach to likelihood-free inference is the synthetic likelihood method [362, 115]. [362] considers the problem of statistical inference of chaotic dynamical systems in ecology. There the likelihood of interest is given by the marginal of a (parametrised) joint distribution over the observables and the noise, and the system is chaotic in the sense that "small" changes in the noise imply (unpredictable) "large" changes in the joint distribution. The consequence is that standard Bayesian inference techniques cannot be applied. As a way forward, the author proposes to first determine which properties of the observables are scientifically meaningful. This defines a "statistic", i.e., a function of the observables, which is assumed to be close to a Gaussian distribution. Then a statistical manifold of synthetic Gaussian distributions (5.1) is constructed, where the mean and covariance associated to a particular parameter θ is determined numerically by simulating the dynamics. Once the synthetic model is in place, the optimal parameter is estimated using MLE. For complex models determining

meaningful summary statistics which are indeed approximately Gaussian may be very challenging. It is thus of interest to develop other likelihood-free methods which are not constrained by distributional assumptions. Similarly, approximate Bayesian Computation (ABC) have also been developed. For a fixed set of observations, these involve sampling a parameter, sampling a new observation given the parameter, calculate the distance with respect to some metric between an appropriate statistic (function) of the new observation and the statistic of the fixed observation. If the distance is smaller than a fixed "tolerance value", the parameter is accepted. These methods generally yield biased samples from the posterior (unless the tolerance is set to zero) but permits likelihood-free inference [37, 247, 248, 141].

We introduced in §A.4.6 an example of such likelihood-free inference based on generative models. Recall that given a probability space $(\mathcal{U}, \mathbb{U})$, we defined a **generative model** as a statistical model whose distributions can be written as the pushforward $\mathbb{P}_\theta \equiv G_\theta^{\#}\mathbb{U}$ of the distribution $\mathbb{U}$ by a generator, that is a family of maps $G_\theta : \mathcal{U} \to \mathcal{X}$. It follows that to generate a sample from $\mathbb{P}_\theta$ that if $U \sim \mathbb{U}$, then $G_\theta \circ U \sim \mathbb{P}_\theta$ (since the pushforward of the composition of maps is the composition of the pushforwards), and thus if we produce i.i.d. realisations $\{u_i\}_{i=1}^n$ then $x_i \equiv G_\theta(u_i)$ will be a i.i.d. realisation from $\mathbb{P}_\theta$. Hence if it straightforward to generate observations from $\mathbb{U}$, and the generator is cheap to evaluate, we can easily produce observations from generative models. Yet, a likelihood function need not be available. We therefore need to construct alternatives to the MLE, as well as the DKSD and DSM estimators discussed in §6. We shall here investigate estimators constructed by minimising approximations to the maximum mean discrepancy.

A.8.1 The Maximum Mean Discrepancy Statistical Manifold

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a characteristic scalar kernel on $\mathcal{X} \subset \mathbb{R}^d$. Given a (not necessarily smooth) target distribution $\mathbb{Q}$, and a parametrised statistical manifold $\mathcal{P}_\Theta(\mathcal{X})$, with Θ an open subset of $\mathbb{R}^p$, we consider the minimum estimators of the MMD^2 statistical divergence (5.7):

$$\hat{\theta}_m = \arg\min_{\theta\in\Theta} \mathrm{MMD}^2(\mathbb{P}_\theta \| \mathbb{Q}^m), \tag{A.16}$$

where $\mathbb{Q}^m(\mathrm{d}y) = \frac{1}{m}\sum_{i=1}^m \delta_{y_i}(\mathrm{d}y)$, and $\{y_i\}_{i=1}^m \overset{\text{IID}}{\sim} \mathbb{Q}$. In the following we will use $\mathbb{Q}^m$ to denote both the random measure $\mathbb{Q}^m$ and the measure $\mathbb{Q}^m(\omega)$, and we shall assume that $\mathcal{P}_\Theta(\mathcal{X}) \subset \mathcal{P}_k(\mathcal{X})$, i.e., the kernel is Bochner integrable with respect to the model distributions. Of course, in general, the optimisation problem will not be convex, and we will not have a tractable expression for the minimiser $\hat{\theta}_m$. If the generator $\theta \mapsto G_\theta$ is differentiable with a tractable derivative, the minimiser will be a fixed point of the equation $\dot{\theta} = -\nabla_\theta \mathrm{MMD}^2(\mathbb{P}_\theta || \mathbb{Q}^m)$ where $\nabla_\theta = (\partial_{\theta_1}, \ldots, \partial_{\theta_p})$. Assuming that the Jacobian $\nabla_\theta G_\theta$ is $\mathbb{U}$-integrable then the gradient term

can be written as

$$\nabla_\theta \mathrm{MMD}^2(\mathbb{P}_\theta \| \mathbb{Q}^m) = 2 \int_{\mathcal{U}} \int_{\mathcal{U}} \nabla_1 k(G_\theta(u), G_\theta(v)) \nabla_\theta G_\theta(u) \mathbb{U}(\mathrm{d}u) \mathbb{U}(\mathrm{d}v) - \frac{2}{m} \sum_{j=1}^{m} \int_{\mathcal{U}} \nabla_1 k(G_\theta(u), y_j) \nabla_\theta G_\theta(u) \mathbb{U}(\mathrm{d}u),$$

where $\nabla_1 k$ corresponds to the partial derivative with respect to the first argument. Since in practice it will not be possible to compute the integral, we introduce a U-statistic approximation for the gradient as follows:

$$\hat{J}_\theta(\mathbb{Q}^m) = \frac{2 \sum_{i \neq i'} \nabla_\theta G_\theta(u_i) \nabla_1 k(G_\theta(u_i), G_\theta(u_{i'}))}{n(n-1)} - \frac{2 \sum_{j=1}^{m} \sum_{i=1}^{n} \nabla_\theta G_\theta(u_i) \nabla_1 k(G_\theta(u_i), y_j)}{nm},$$

where $\{u_i\}_{i=1}^n \overset{\text{IID}}{\sim} \mathbb{U}$. This is an unbiased estimator in the sense that $\mathbb{E}[\hat{J}_\theta(\mathbb{Q}^m)] = \nabla_\theta \mathrm{MMD}^2(\mathbb{P}_\theta || \mathbb{Q}^m)$, where the expectation is taken over the independent realisations of the $u_i's$. Hence we can apply a Euclidean stochastic gradient descent algorithm: starting from $\hat{\theta}^{(0)} \in \Theta$:

(i) Sample $\{u_i\}_{i=1}^n \overset{\text{IID}}{\sim} \mathbb{U}$ and compute $x_i \leftarrow G_{\hat{\theta}^{(k-1)}}(u_i)$ for $i = 1, \ldots, n$.

(ii) $\hat{\theta}^{(k)} \leftarrow \hat{\theta}^{(k-1)} - \eta_k \hat{J}_{\hat{\theta}^{(k-1)}}(\mathbb{Q}^m)$.

We also introduce the minimum unbiased estimators $\hat{\theta}_{n,m} = \mathrm{argmin}_{\theta \in \Theta} \mathrm{MMD}^2_{U,U}(\mathbb{P}^n_\theta \| \mathbb{Q}^m)$ of the MMD^2 statistical divergence, with

$$\mathrm{MMD}^2_{U,U}(\mathbb{P}^n_\theta \| \mathbb{Q}^m) = \frac{\sum_{i \neq i'} k(x_i, x_{i'})}{n(n-1)} - \frac{2 \sum_{j=1}^{m} \sum_{i=1}^{n} k(x_i, y_j)}{mn} + \frac{\sum_{j \neq j'} k(y_j, y_{j'})}{m(m-1)}.$$

where $\mathbb{P}^n_\theta = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$ for some $\{x_i\}_{i=1}^n \overset{\text{IID}}{\sim} \mathbb{P}_\theta$, which provides useful insights in the cases where the generator gradient $\nabla_\theta G_\theta$ is computationally intensive.

A.8.2 The Natural Gradient Flow induced by MMD

Rather than using Euclidean gradient descent, we found in our experiments that is preferable to employ the metric tensor on Θ induced by MMD, as we now discuss. Assuming that the map $\theta \to \mathbb{P}_\theta$ is injective, the MMD distance between the elements $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ in $\mathcal{P}_k$ induces a distance between θ and θ' in Θ. In this section, we study instead the geometry induced by MMD. We assume that (i) $G_\theta(\cdot)$ is -measurable for all $\theta \in \Theta$; (ii) $G_\cdot(u) \in C^1(\Theta)$ for all $u \in \mathcal{U}$; (iii) $\|\nabla_\theta G_\theta(\cdot)\| \in L^1(\mathbb{U})$, for all $\theta \in \Theta$. Suppose additionally that the kernel k has bounded continuous derivatives over $\mathcal{X} \times \mathcal{X}$. Define the map $J : \Theta \to \mathcal{H}_k$ to be the Bocher integral $J(\theta) = \Pi_k(\mathbb{P}_\theta)$. By [149, Theorem 90], assumptions (i)-(iii) imply that the map J is Fréchet

differentiable and

$$\partial_{\theta_i} J(\theta)(\cdot) = \int_{\mathcal{U}} \nabla_2 k(\,\cdot\,, G_\theta(u)) \partial_{\theta_i} G_\theta(u) \mathbb{U}(\mathrm{d}u).$$

The map J induces a degenerate-Riemannian metric $g(\theta)$ on Θ given by the pull-back of the inner product on $\mathcal{H}_k$. In particular its components in the local coordinate-system are $g_{ij}(\theta) = \langle \partial_{\theta_i} J(\theta), \partial_{\theta_j} J(\theta) \rangle_{\mathcal{H}_k}$ for $i, j \in \{1, \ldots, p\}$. By [328, Lemma 4.34], it follows that for $i, j \in \{1, \ldots, p\}$,

$$g(\theta) = \int_{\mathcal{U}} \int_{\mathcal{U}} \nabla_\theta G_\theta(u)^\top \nabla_2 \nabla_1 k(G_\theta(u), G_\theta(v)) \nabla_\theta G_\theta(v) \mathbb{U}(\mathrm{d}u) \mathbb{U}(\mathrm{d}v), \tag{A.17}$$

where $\nabla_1 \nabla_2 k(x, y) = \{\partial_{x_i} \partial_{y_j} k(x, y)\}_{i,j=1,\ldots,d}$. The induced metric tensor is in fact just the information metric associated to the MMD-squared statistical divergence (see §5.2).

As usual we typically do not have a tractable expression for g above, and will thus use an unbiased approximation

$$g_U(\theta) = \frac{1}{n(n-1)} \sum_{i \neq j} \nabla_\theta G_\theta(u_i)^\top \nabla_2 \nabla_1 k(G_\theta(u_i), G_\theta(u_j)) \nabla_\theta G_\theta(u_j),$$

where $\{u_i\}_{i=1}^n$ are IID realisations from $\mathbb{U}$. Hence, we thus replace the Euclidean stochastic gradient descent algorithm introduced in the previous section with the following natural stochastic gradient descent algorithm (see (5.3)): given $\hat{\theta}^{(0)} \in \Theta$:

(i) Sample $\{u_i\}_{i=1}^n \overset{\text{IID}}{\sim} \mathbb{U}$ and set $x_i \leftarrow G_{\hat{\theta}^{(k-1)}}(u_i)$ for $i = 1, \ldots, n$.

(ii) Compute $\hat{\theta}^{(k)} \leftarrow \hat{\theta}^{(k-1)} - \eta_k g_U\big(\hat{\theta}^{(k-1)}\big)^{-1} \hat{J}_{\hat{\theta}^{(k-1)}}(\mathbb{Q}^m)$.

A.8.3 Theory of Minimum MMD estimators

We now summarise some useful theoretical guarantees for the estimators above which I proved together with Andrew Duncan. The details of the proofs may be found in [62], which also contains many experiments showing these estimators perform well.

A.8.3.1 Concentration and Generalisation Bounds for MMD

We will make the following standard assumptions:

Assumption 1. *1. For every $\mathbb{Q} \in \mathcal{P}_k(\mathcal{X})$, there exists $c > 0$ such that the set $\{\theta \in \Theta : MMD(\mathbb{P}_\theta || \mathbb{Q}) \leq \inf_{\theta' \in \Theta} MMD(\mathbb{P}_{\theta'} || \mathbb{Q}) + c\}$, is bounded.*

2. *For every* $n \in \mathbb{N}$ *and* $\mathbb{Q} \in \mathcal{P}_k(\mathcal{X})$*, there exists* $c_n > 0$ *such that the set* $\{\theta \in \Theta : MMD(\mathbb{P}_\theta^n||\mathbb{Q}) \leq \inf_{\theta' \in \Theta} MMD(\mathbb{P}_{\theta'}||\mathbb{Q}) + c_n\}$*, is almost surely bounded.*

Theorem 65 (Generalisation Bounds)**.** *Suppose that the kernel* k *is bounded, and that Assumption 1 holds, then with probability at least* $1 - \delta$*,*

$$MMD\left(\mathbb{P}_{\hat{\theta}_m}||\mathbb{Q}\right) \leq \inf_{\theta \in \Theta} MMD(\mathbb{P}_\theta||\mathbb{Q}) + 2\sqrt{\frac{2}{m}\sup_{x \in \mathcal{X}} k(x,x)}\left(2 + \sqrt{\log\left(\frac{1}{\delta}\right)}\right),$$

and

$$MMD\left(\mathbb{P}_{\hat{\theta}_{n,m}}||\mathbb{Q}\right) \leq \inf_{\theta \in \Theta} MMD(\mathbb{P}_\theta||\mathbb{Q}) + 2\left(\sqrt{\frac{2}{n}} + \sqrt{\frac{2}{m}}\right)\sqrt{\sup_{x \in \mathcal{X}} k(x,x)}\left(2 + \sqrt{\log\left(\frac{2}{\delta}\right)}\right).$$

It follows that the rate of convergence decreases as $n^{-\frac{1}{2}}$ and $m^{-\frac{1}{2}}$, independently of the dimensions p and d. A similar generalisation result was obtained in [106] for minimum MMD estimation of deep neural network models. While the bounds are of the same form, Theorem 65 only requires minimal assumptions on the smoothness of the kernel. Moreover, all the constants in the bound are explicit, demonstrating clearly dimensional dependence. We also have the following concentration inequality.

Lemma 66 (Concentration Bound)**.** *Assume that the kernel* k *is bounded. Let* $\mathbb{P}^n$ *be the empirical measure obtained from* n *independently* $\mathbb{P}$*-distributed random variables. Then with probability* $1 - \delta$*, we have that*

$$MMD(\mathbb{P}||\mathbb{P}^n) \leq \sqrt{\frac{2}{n}\sup_{x \in \mathcal{X}} k(x,x)}\left(1 + \sqrt{\log\left(\frac{1}{\delta}\right)}\right).$$

A.8.3.2 Consistency and Asymptotic Normality

We have the following strong consistency result for our estimators:

Proposition 67 (Consistency)**.** *Suppose that Assumption 1 holds and that there exists a unique minimiser* $\theta^* \in \Theta$ *such that* $MMD(\mathbb{P}_{\theta^*}||\mathbb{Q}) = \inf_{\theta \in \Theta} MMD(\mathbb{P}_\theta||\mathbb{Q})$*. Then* $\lim_{m \to \infty} \hat{\theta}_m = \theta^*$ *and* $\lim_{m,n \to \infty} \hat{\theta}_{m,n} = \theta^*$ *as* $n, m \to \infty$*, almost surely.*

Assuming consistency, we furthermore have the following central limit theorem in the M-close setting, i.e. assuming that $\mathbb{Q} = \mathbb{P}_{\theta^*}$ for some θ^*.

Theorem 68 (Central Limit Theorems)**.** *Suppose that* $\mathbb{Q} = \mathbb{P}_{\theta^*}$ *for some* $\theta^* \in \Theta$ *and that the conclusions of Proposition 67 hold. Suppose that:*

1. *There exists an open neighbourhood* $O \subset \Theta$ *of* θ^* *such that* G_θ *is three times differentiable in* O *with respect to* θ.

2. *The information metric* $g(\theta)$ *is positive definite at* $\theta = \theta^*$.

3. *There exists a compact neighbourhood* $K \subset O$ *of* θ^* *such that* $\int_{\mathcal{U}} \sup_{\theta \in K} \left\| \nabla^{(i)} G_\theta(u) \right\| \mathbb{U}(\mathrm{d}u) < \infty$ *for* $i = 1, 2, 3$ *where* $\nabla^{(i)}$ *denotes the mixed derivatives of order* i *and* $\|\cdot\|$ *denotes the spectral norm.*

4. *The kernel* $k(\cdot, \cdot)$ *is translation invariant, with bounded mixed derivatives up to order* 2.

Then as $k \to \infty$*:*

$$\sqrt{m}\left(\hat{\theta}_m - \theta^*\right) \xrightarrow{d} \mathcal{N}(0, C),$$

where $\xrightarrow{d}$ *denotes convergence in distribution. The covariance matrix is given by* $C = g(\theta^*)^{-1} \Sigma g(\theta^*)^{-1}$ *where*

$$\Sigma = \int_{\mathcal{U}} \left(\int_{\mathcal{U}} \left(\nabla_1 k(G_{\theta^*}(u), G_{\theta^*}(v)) \nabla_\theta G_{\theta^*}(u) - \overline{M} \right) \mathbb{U}(\mathrm{d}u) \right)^{\otimes 2} \mathbb{U}(\mathrm{d}v)$$

and

$$\overline{M} = \int_{\mathcal{U}} \int_{\mathcal{U}} \nabla_1 k(G_{\theta^*}(u), G_{\theta^*}(v)) \nabla_\theta G_{\theta^*}(u) \mathbb{U}(\mathrm{d}u) \mathbb{U}(\mathrm{d}v).$$

Here, $A^{\otimes 2} := A \otimes A$. *Furthermore, suppose that:*

5 *The kernel* $k(\cdot, \cdot)$ *has bounded mixed derivatives up to order* 3.

6 *The indices satisfy* $n = n_k$, $m = m_k$ *where* $n_k/(n_k + m_k) \to \lambda \in (0, 1)$,

Then, as $k \to \infty$,

$$\sqrt{n_k + m_k}\left(\hat{\theta}_{n,m} - \theta^*\right) \xrightarrow{d} \mathcal{N}(0, C_\lambda),$$

where $C_\lambda = (1/(1-\lambda)\lambda)C$.

To the best of our knowledge, there are no known analogous result for minimum Wasserstein or Sinkhorn estimators (except in one dimensions).

Recall from §2.2 that he Cramer-Rao bound provides a lower bound on the variance of any unbiased estimator for $\mathbb{P}_\theta$. The following result adapts the bound to our biased estimators.

Theorem 69 (Cramer-Rao Bounds). *Suppose that the CLTs in Theorem 68 hold and that the data distribution $\mathbb{Q}$ satisfies $\mathbb{Q} = \mathbb{P}_{\theta^*}$, where $\mathbb{P}_{\theta^*} = G_{\theta^*}^{\#}\mathbb{U}$ is assumed to have density $p(x|\theta^*)$. Furthermore, suppose that the MMD information metric $g(\theta^*)$ and the Fisher information metric $F(\theta) = \int_{\mathcal{X}} \nabla_\theta \log p(x|\theta) \nabla_\theta \log p(x|\theta)^\top \mathbb{P}_\theta(\mathrm{d}x)$ are positive definite when $\theta = \theta^*$. Then the asymptotic covariances C and C_λ of the estimator $\hat{\theta}_m$ and $\hat{\theta}_{n,m}$ satisfy the Cramer-Rao bound, i.e. $C - F(\theta^*)^{-1}$ and $C_\lambda - F(\theta^*)^{-1}$ are non-negative definite.*

A.8.3.3 Robustness

Finally we can analyse the robustness of our estimators using the influence function as in §6.7. To the best of our knowledge, the robustness of Wasserstein and Sinkhorn estimators has not been studied.

Theorem 70 (Bias Robustness). *The influence function corresponding to the maximum mean discrepancy is given by $IF_{MMD}(z, \mathbb{P}_\theta) = g^{-1}(\theta) \nabla_\theta MMD(\mathbb{P}_\theta, \delta_z)$. Furthermore, suppose that $\nabla_1 k$ is bounded and $\int_{\mathcal{U}} \|\nabla_\theta G_\theta(u)\| \mathbb{U}(\mathrm{d}u) < \infty$, then the MMD estimators are bias-robust.*