Simulation-based inference for astrophysical data

Anau Montel, N.

Publication date
2024
Document Version
Final published version

Simulation-based inference for astrophysical data

Noemi Anau Montel

Simulation-based inference for astrophysical data

Noemi Anau Montel

Simulation-based inference for astrophysical data

This work has been accomplished at the Gravitation and AstroParticle Physics Amsterdam (GRAPPA) centre of excellence and the Institute for Theoretical Physics (ITFA) of the University of Amsterdam (UvA). This research is funded by the European Research Council (ERC) Consolidator Grant No. 864035 UnDark (PI Weniger).

Simulation-based inference for astrophysical data

Academisch Proefschrift

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

prof. dr. ir. P.P.C.C. Verbeek

ten overstaan van een door het College voor Promoties

ingestelde commissie,

in het openbaar te verdedigen in de Agnietenkapel

op woensdag 30 oktober 2024, te 10:00 uur

door

Noemi Anau Montel

geboren te Torino

Promotiecommissie

Promotores

dr. C. Weniger	Universiteit van Amsterdam
dr. S. Ando	Universiteit van Amsterdam

Overige leden

prof. dr. G. Bertone	Universiteit van Amsterdam
prof. dr. F. Donato	University of Torino
prof. dr. N. Fornengo	University of Torino
dr. S. Nissanke	Universiteit van Amsterdam
dr. S. Vegetti	Max Planck Institute for Astrophysics
dr. J. Vink	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Publications

This thesis is based on the following publications:

[1] N. Anau Montel, A. Coogan, C. Correa, K. Karchev, C. Weniger, *Estimating the warm dark matter mass from strong lensing images with truncated marginal neural ratio estimation*, Mon. Not. Roy. Astron. Soc. 527 (2024) 66, arXiv:2209.09918.

All authors participated in the planning of the project. The analysis code was primarily written by NAM with input from AC, KK, and CW. The paper was primarly written by NAM with inputs from all authors.

Presented in Chapter 3.

[2] N. Anau Montel, J. Alvey, C. Weniger, *Scalable inference with Autoregressive Neural Ratio Estimation*, Mon. Not. Roy. Astron. Soc. 527 (2024) 66, arXiv:2308.08597.

All authors participated in the planning of the project. The analysis code was primarily written by NAM, except for the stellar stream example contributed by JA. The paper was written by NAM and JA, with input from CW.

Presented in Chapter 4.

[3] F. List, N. Anau Montel, C. Weniger, *Bayesian Simulation-based Inference for Cosmological Initial Conditions*, Machine Learning and the Physical Sciences Workshop at the 37th Conference on Neural Information Processing Systems (NeurIPS 2023) [Paper] [Poster] , arXiv:2310.19910.

All authors participated in the planning of the project. FL provided the simulations, NAM and FL wrote the analysis code with input from CW. The paper was written by both NAM and FL, with input from CW.

Presented in Chapter 5.

[4] N. Anau Montel, C. Weniger, *Detection is truncation: studying source populations with truncated marginal neural ratio estimation*, Machine Learning and the Physical Sciences Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022) [Paper] [Poster] , arXiv:2211.04291.

Both authors participated in the planning of the project. The analysis code was written by NAM with input from CW. The paper was primarly written by NAM with input from CW.

Presented in Chapter 6.

Other publications by the author:

[5] A. Coogan, N. Anau Montel, K. Karchev, M. W. Grootes, F. Nattino, C. Weniger, *The effect of the perturber population on subhalo measurements in strong gravitational lenses*, Mon. Not. Roy. Astron. Soc. 518 (2023) 2746, arXiv:2205.09126.

[6] K. Karchev, N. Anau Montel, A. Coogan, C. Weniger, *Strong-Lensing Source Reconstruction with Denoising Diffusion Restoration Models*, Machine Learning and the Physical Sciences Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022) [Paper] [Poster], arXiv:2211.04365.

[7] C. Correa, M. Schaller, S. Ploeckinger, N. Anau Montel, C. Weniger, S. Ando, *TangoSIDM: Tantalizing models of Self-Interacting Dark Matter*, Mon. Not. Roy. Astron. Soc. 517 (2022) 3045, arXiv:2206.11298.

Contents

1 Introduction

The fundamental goal of the physical sciences is to understand the laws that govern physical phenomena. This process inherently involves a continuous interplay between theoretical models and physical observations, forming a feedback loop that continuously refines both. By comparing observational data with model predictions, model validity can be tested, areas of improvements can be identified, and their faithfulness to reality can be improved. Conversely, theoretical models can guide observational strategies by predicting phenomena that have yet to be observed, hence focusing the attention to specific phenomena or conditions that may yield new insights.

A key ingredient of this scientific process is *statistics*. Statistics is a rigorous mathematical language that enables formal statements about what events are possible under physical laws, thus bridging the gap between physics models and observational data. Statistics is today at the very heart of the scientific process, and not just an optional nuisance. Guided by statistical formulation, given some parameters that describe a physical system, implied *predictions* or consequences of a physical model can be computed, allowing for the systematic exploration of the model's implications. Predictions can be either deterministic or intrinsically stochastic, e.g. due to the randomness of the physical processes, the measurement processes, or incomplete information. In order to refine theoretical models, it is essential to perform the inverse process: starting with the effects to discover the causes, inferring from a set of observations the causal factors that produced them. This task, known as solving an *inverse problem* [8, 9], involves mapping back observational data to infer the underlying model parameters that are not directly observable.

In astrophysics, this iterative cycle between prediction and inference is particularly challenging due to the complex nature of the systems under study. In this field, countless statistical methods have been developed over the years, and although rarely in the spotlight, they ultimately determine what is accepted as scientifically established truth. Since an inefficient or incorrect use of statistical data analysis may lead to weaker or entirely wrong conclusions, it is thus of the utmost importance to identify where and how progress is possible in order to advance the field.

1.1 Astrophysical data analysis challenges

We are at the dawn of a data-driven era in astrophysics and cosmology. As we can see from Figure 1.1, that shows the minimum volume of data per year expected to be produced by a range of recent and upcoming surveys and experiments, the coming decade will see transformative science conducted by observatories, based both on the ground (e.g. Rubin-LSST [10], ELT [11]), and space-based missions (e.g. JWST [12], Euclid [13]). Crucially, this wealth of data promises unprecedented high-precision measurements of the growth of structure and geometry of the universe, opening new windows to dark matter [14, 15], dark energy [16, 17], neutrino physics [18, 19], and inflationary cosmology [20, 21].

Given the unprecedented size and detail of these data, connecting theoretical models with this wealth of high-precision observations presents significant challenges, and the scientific return of many upcoming observations is expected to be limited by the efficiency of our statistical inference tools [22, 23]. First, information must be optimally extracted from the data to avoid discarding valuable insights. Second, uncertainties must be correctly treated and thoroughly propagated to ensure accurate scientific statements. Moreover, fully exploiting this data for scientific purposes will require increasingly complex physical models, which bring along higher computational costs, as well as a larger number of uncertain parameters, including those characterizing signal and background systematics. Hence, the need for principled and scalable statistical analyses has never been more critical (for recent reviews on statistical analysis developments in cosmology and astrophysics see Refs. [24–26]).

To get a glimpse of the types of challenges astrophysical data analysis is currently facing, we will now discuss on a high-level some exemplary problems. The selected topics cover only a small part of the current astrophysical statistical analysis challenges, and reflect the works presented in this thesis.

Strong gravitational lensing Gravitational lensing [27], the phenomenon where light-rays from a distant source bend due to the presence of an intervening massive object (the lens), has significantly advanced various fields of physics. Significant findings include, e.g., discovering and analysing some of the most distant galaxies in the universe [e.g. 28–30], determining the dark matter content within galaxy clusters and understanding its distribution on both galactic and sub-galactic scales [e.g. 31–35], detecting individual light dark matter halos [e.g. 36–38], and measuring the Hubble constant [e.g. 39, 40]. Crucial to many of these studies is the capability to accurately invert the physical predictive model for the source and lens. Hence, it includes dealing with very large number of degenerate parameters: for instance, variations in the source morphology (e.g., parametrized by a pixel grid, with $\mathcal{O}$(pixels) parameters [e.g. 41]), substructure population (e.g., tenth of thousands of parameters for a cold dark matter

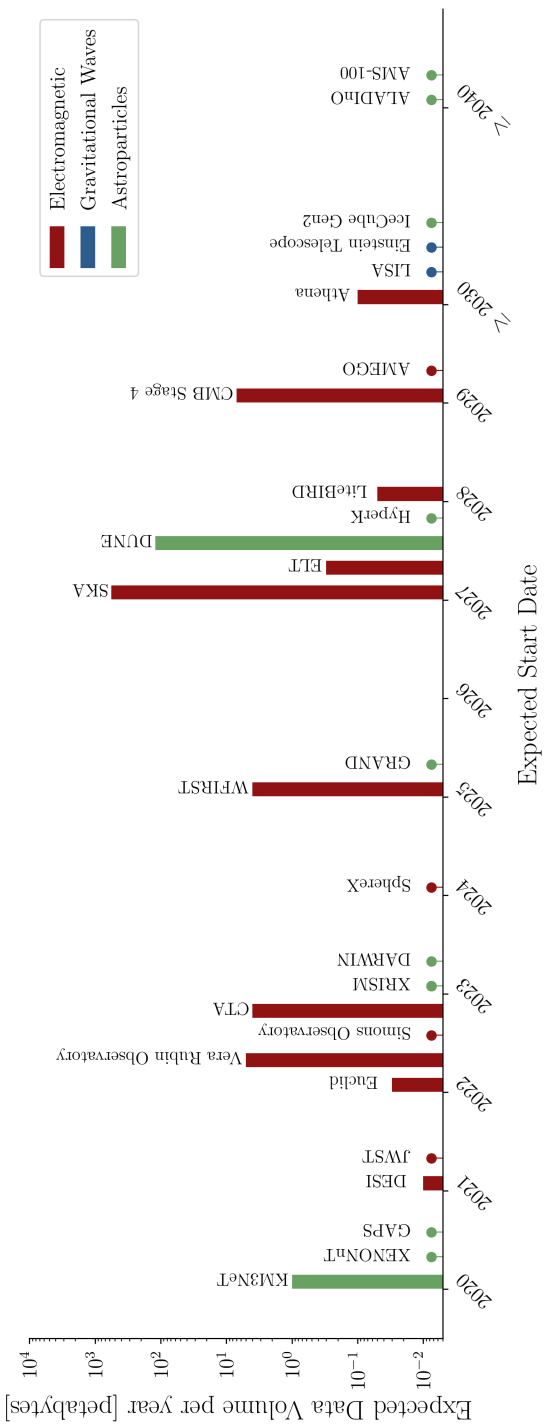

Figure 1.1: The graphic shows rough estimates for the minimum expected volume of data per year in petabytes for a range of current and upcoming experiments and surveys in astrophysics and cosmology. Note that both the data volume and start dates are indicative; the plot is intended to give a general overview of the current and future data analysis challenges of the field. The figure and caption are a reproduction of Figure 12 in Ref. [22].

population [e.g. 1]), and main lens mass distribution. Moreover, it involves solving a mixture of parameter inference, object detection, and image reconstruction problem.

Cosmology Recent and upcoming large-scale structure surveys will measure galaxy distributions with unprecedented precision, improving our understanding of cosmogenesis, neutrino physics, and dark energy [e.g. 10, 42]. These surveys compare cosmological models to data, starting with a model of the universe's initial conditions – an isotropic density field with small Gaussian perturbations. This field evolved through gravitational processes into today's structures, catalogued in redshift space [43]. By modeling early universe processes and understanding the evolution of perturbations influenced by dark matter and dark energy, large-scale structure surveys provide direct constraints on the initial density field and its evolution [44]. Mapping back today's structures to constraint the initial density field requires developing advanced cosmological simulators, and accounting for non-linear transformations, mixing of spatial structure, and noise, all while exploring a multi-million-dimensional parameter space at the field level.

Point-source in sky maps A striking example of how having higher-resolution data calls for more complex models that need to account for more objects, hence more parameters, hence harder inference, is given by the evolution of γ-ray observations in the galactic plane. Starting with SAS-2 satellite (1972-1073), 6 γ-ray sources and the diffuse emission were identified [45, 46]. Two decades later, with EGRET (1991-2000), it was possible to identify up to 188 γ-ray sources, extended point sources, and resolve gas emission [47]. With Fermi-LAT [48], we have now more than 6600 γ-ray sources, highly resolved gas maps, many extended sources, the detection of the Fermi bubbles [49], and an unidentified component, the so called GeV excess [50]. A physical model to describe this type of data can require, e.g., $\mathcal{O}(10^4)$ parameters for the resolved point sources alone. Moreover, some of its components have an unclear uncertainty quantification (e.g. the gas maps). For this type of data, the statistical analysis can be thought of as a mixture of source detection and image analysis, and requires self-consistent measurement of point-source population parameters based on both detected and undetected objects.

The common thread behind these physical systems and their statistical challenges is that they are formally representable by large models, with the adjective "large" describing three different properties at the same time: their complexity (in terms of number of components and potentially intricate interactions between them), their volume (in terms of observational data), and the computational resources (in terms of power, memory, and time) needed to solve them. In recent years, new classes of scalable, fast, and computational-efficient inference algorithms have been enabled by breakthroughs in machine learning [51], which could play a major role in the successful analysis of these types of astrophysical data.

1.2 The emergence of the simulation-based inference paradigm

Over the years, in order to draw scientific conclusions, an abundance of statistical tools and physics simulation codes have been developed within the community.[1] In particular, *physics simulators* have always served as powerful predictive devices, mapping model parameters into realized data by reproducing numerically the underlying natural phenomenon of interest. Recently, remarkable progress in computing technologies and programming languages have made it possible to express increasingly detailed and complex physical models through high-fidelity computer simulators. However, while simulators excel at predicting system behaviors, they are poorly suited for statistical inference and for solving inverse problems. Broadly speaking, to evaluate the likelihood of a data realization implicitly defined through a computer simulator one must solve an inverse problem that involves integrating all possible code paths, for all possible simulator configurations, that could have potentially led to the observed data realization. Clearly, as the fidelity and detail of modern computer simulations increase, computing this quantity becomes exceedingly difficult, if not entirely intractable or computationally infeasible [52].

Fortunately, recent advances in deep learning [53] and differentiable programming [54] have led to the emergence and proliferation of a *simulation-based inference paradigm* that can effectively tackle the above challenges. By leveraging the power of neural networks, these new methods can approximate the complex relationships within simulators, allowing for efficient solutions to inverse problems that were previously beyond reach [52]. This paradigm shift in statistical analysis has led to the proliferation of tools for simulation-based inference [e.g. 55–58] and to their application to various problems in gravitational waves astronomy [e.g. 59–66], strong gravitational lensing analysis [e.g. 1, 5, 67–71], cosmological probes [e.g. 3, 55, 72–82], and many other astrophysical problems [e.g. 4, 83–92].

We will see more in depth in Chapter 2 the technical details of the various simulation-based inference algorithms, showing how they can improve the quality of insight we can gain from simulations, maximizing information extraction from data, and providing robust uncertainty quantification for scientific statements. In the next introductory section, we will focus on providing the context and motivation for this paradigm shift, comparing it to the likelihood-based paradigm, and highlighting how it can help in tackling some of the challenges discussed previously in Section 1.1.

[1] An extensive list of the tools used in astrophysics, cosmology, and high energy physics can be found here: https://github.com/nikosarcevic/HEP-ASTRO-COSMO.

1.3 To likelihood-base or to simulation-base?

Ultimately, we are interested in inferring the probability distribution of model parameters $\boldsymbol{\Theta}$ for a given observation $\boldsymbol{x}_0$. In a Bayesian inference context, the posterior distribution for model parameters $\boldsymbol{\Theta}$ follows from Bayes' theorem

$$p(\boldsymbol{\Theta} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid \boldsymbol{\Theta})}{p(\boldsymbol{x})} p(\boldsymbol{\Theta}) \,, \tag{1.1}$$

where $p(\boldsymbol{x} \mid \boldsymbol{\Theta})$ is the likelihood of the data $\boldsymbol{x}$ for given parameters $\boldsymbol{\Theta} \in \mathbb{R}^D$, $p(\boldsymbol{\Theta})$ is the prior probability distribution over the parameters, and $p(\boldsymbol{x})$ is the evidence of the data. As evident from Equation (1.1), the Bayesian framework needs both a formalization of the modeling assumptions, encoded by the likelihood, and a prior knowledge associated with each learnable parameter of the model, encoded by the prior.

Given this Bayesian setup, statistical inference is performed within the context of a probabilistic model $p(\boldsymbol{x} \mid \boldsymbol{\Theta})$, that can be in principle accessed by two different routes. On one hand, likelihood-based inference (LBI) algorithms rely on likelihood evaluations, single scalars that quantify closeness to the observation $\boldsymbol{x}_0$. On the other hand, SBI algorithms do not explicitly calculate the likelihood function, but instead rely on samples from a stochastic simulator that *implicitly maps* model parameters $\boldsymbol{\Theta}$ to data $\boldsymbol{x}$.

Likelihood-based methods

The main LBI tools to solve inverse problems for modern astrophysical and cosmological data analysis have been sampling-based inference methods, like Markov-chain Monte Carlo (MCMC) [93, 94] and nested sampling [95–97] techniques. However, it is especially challenging to ensure these methods convergence in high dimensional parameter spaces (the time needed to reach convergence scales poorly with the dimensionality of the explored parameter space), for multi-modal posteriors, and curving degeneracies. More modern methods are taking up these challenges, including Gibbs samplers [98] (that rely on conditional distributions), and slice-sampling techniques [99, 100]. For example, the new generation nested sampler `PolyChord` [100], based on slice-sampling, has at worst a $\mathcal{O}(D^3)$ scaling, whereas `Multinest` [96] has an exponential scaling that emerges at high dimensions (see Figure 4 in Ref. [100]).

As the dimensionality grows, sampling from the typical set of the posterior distribution becomes exponentially difficult [101]. Therefore, it is useful to resort to gradient-based algorithms, as they are able to concentrate the sampling in high posterior mass regions, despite the large number of model parameters, provided one has efficient access to accurate derivatives of the likelihood function with respect to the

model parameters [101]. The most popular gradient-based method among physicist is Hamiltonian Monte Carlo (HMC) [102, 103], which is built on the formalism of Hamiltonian dynamics (as the name implies). Widely used in cosmology, it is of particular notice the application of Ref. [44]: dynamical large-scale structure inference from galaxy redshift surveys, where the parameter space is the 3D initial matter density field, of order $\mathcal{O}(10^7)$ voxels. This impressive result has been achieved with significant computational resources and by carefully tuning the so called "mass matrix" of the HMC, which is not always possible.

Lastly, it is worth mentioning variational inference (VI), which, differently from the previously mentioned LBI methods, allows for the approximation of extremely high-dimensional Bayesian posteriors with simple proposal distributions by solving an optimization problem [104, 105]. As for HMC, VI's efficient implementation requires gradients from an end-to-end differentiable physical simulator [e.g. 106–108]. This can be achieved with little extra effort through auto-differentiation libraries that are standard in deep learning packages [e.g. 109, 110]. Notably, VI has already been successfully applied in various astrophysical contexts [41, 111–114]. Of particular notice is Ref. [111], where they jointly optimize parameters for 188×10^6 stars and galaxies using tera-scale datasets. We will come back to VI in Section 2.1.2, in order to compare it in details with its direct SBI counterpart.

Simulation-based methods

Differently from LBI methods, SBI entirely relies on samples from a stochastic simulator that maps model parameters $\boldsymbol{\Theta}$ to data $\boldsymbol{x}$. This mapping is equivalent to sampling from the model distribution $\boldsymbol{x} \sim p(\boldsymbol{x} \mid \boldsymbol{\Theta})$, which is effectively an implicit representation of the likelihood. As a result, in this setting one just needs a computational code that generates random samples from $p(\boldsymbol{x} \mid \boldsymbol{\Theta})$, that can be later used by a SBI algorithm. For the purpose of this thesis, a *simulator/forward model* is a computer program that takes as input a vector of parameters $\boldsymbol{\Theta} \in \mathbb{R}^D$, samples a series of internal states or latent variables, and finally produces a data vector $\boldsymbol{x}$ as output (usually our observable). Programs that involve random samplings and are interpreted as statistical models are known as probabilistic programs, and simulators are an example [52]. In principle, using simulators allows for the simultaneous inclusion of all relevant processes that can affect the data, regardless of whether a full probabilistic description is tractable or not, as long as they can be efficiently programmed. In this context, *intractability* means one of two things: a closed-form expression of the likelihood distribution is not available, or even if available it is computationally too expensive, e.g., in the worst case, it scales exponentially with the number of parameters [115, 116].

A more detailed exploration of the various SBI algorithms is presented in Chap-

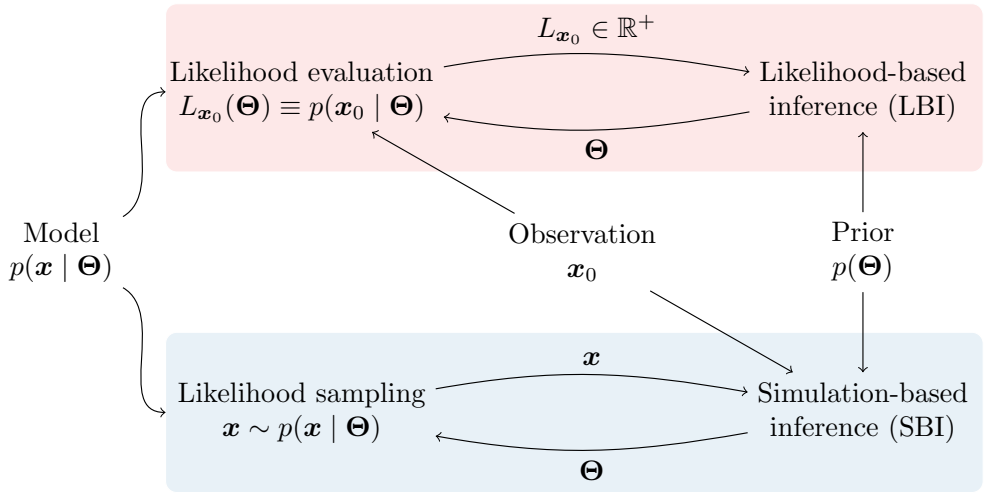

Figure 1.2: *Likelihood-based* inference algorithms rely on the evaluated likelihood $L_{\boldsymbol{x}_0}(\boldsymbol{\Theta})$, which is a single scalar that quantifies closeness to the observation $\boldsymbol{x}_0$. *Simulation-based* inference algorithms learn a function that can be evaluated on many different observations $\boldsymbol{x}_0$, determining their optimal distance measures case by case. Diagram credits: Christoph Weniger.

ter 2, we will now focus on the main high-level differences between LBI and SBI, drawing from the literature to highlight SBI advancements over LBI methods.

Comparison: SBI vs. LBI

The main differences between SBI and LBI methods are summarized in Figure 1.2. In both cases, we start with a data model, $p(\boldsymbol{x} \mid \boldsymbol{\Theta})$, which describes the probability of data $\boldsymbol{x}$ given parameters $\boldsymbol{\Theta}$. In the LBI case, the strategy is a detailed analysis of the likelihood function given an observation $\boldsymbol{x}_0$, $L_{\boldsymbol{x}_0}(\boldsymbol{\Theta}) \equiv p(\boldsymbol{x}_0 \mid \boldsymbol{\Theta})$. To this end, simplifying, the LBI algorithm will suggest points $\boldsymbol{\Theta}$ where the likelihood will be evaluated, and try to focus on regions with high density. On the other hand, SBI techniques do *not* require a tractable (see above) likelihood-density $p(\boldsymbol{x}_0 \mid \boldsymbol{\Theta})$ at a specific observation $\boldsymbol{x}_0$. Instead, they rely on synthetic data samples from the likelihood function $\boldsymbol{x} \sim p(\boldsymbol{x} \mid \boldsymbol{\Theta})$, for a range of model parameters $\boldsymbol{\Theta}$ that are in the simplest case drawn from the parameter priors, $\boldsymbol{\Theta} \sim p(\boldsymbol{\Theta})$, or in more complex cases from generative models [e.g. 6].

While in principle the two frameworks converge to the same answer, when applied several practical differences emerge. We will highlight now two of the most striking disparities, but others will surface in the next chapter when the discussion becomes more technical.

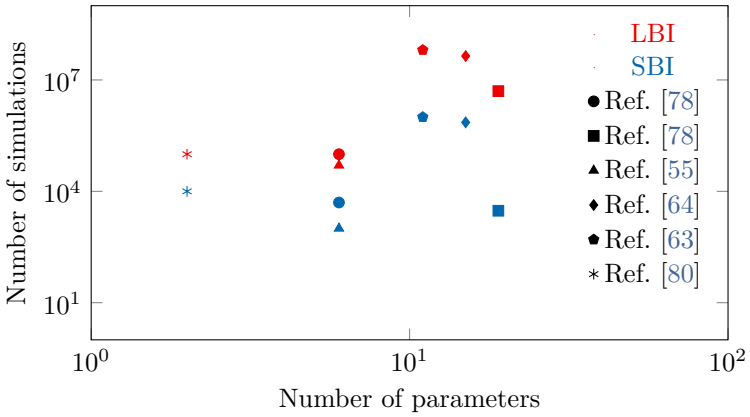

Figure 1.3: Simplified comparison of likelihood-based and simulation-based algorithms in the space of number of required simulations versus number of model parameters. Each work addresses different physics applications, each with its own set of challenges and variations, employing diverse types of analysis to solve them. Therefore, each reported scatter point is unique and achieve the reported scaling thanks to different properties of the various SBI and LBI algorithms employed in the specific study.

Recyclable inference. As highlighted in Figure 1.2, the analyzed observation $\boldsymbol{x}_0$ enters the statistical framework of LBI and SBI at different stages. In particular, LBI algorithms perform inference for a fixed observation $\boldsymbol{x}_0$, and must rerun from scratch for any another observation. It is thus computationally costly to perform new analysis and statistical test on the obtained results. On the other hand, we will see that SBI algorithms effectively learn an estimate of the probability density function that can be used to perform "online" inference on any new data (as long as they stem from the same prior support). In this case, there is no need to rerun the whole pipeline for different observations, but just to re-evaluate the learned function on new data.[2] Furthermore, statistical consistency tests can be performed rather quickly and efficiently. This aspect will be explored in Section 2.3.

Breaking the curse of dimensionality. When using likelihood-based techniques, in order to solve *one* inference problem, like obtaining samples for a marginal posterior of interest, one has to solve *all* of them (joint posterior estimate). The computational overhead of generating joint samples as an intermediate step of marginal inference can be enormous, and can quickly turn an apparently easy inference task, like the measurement of a single physical parameter, into a big challenge. These cases are not uncommon, and, when possible, require problem specific care. For instance, it is

[2]This property is not fully preserved in case of sequential SBI algorithms, as they prioritize achieving other types of benefits (this will be illustrated in Chapter 2).

common in several scenarios to analytically perform parts of the marginal integrals to reduce the parameter space in clever ways: e.g., in cosmic microwave background analysis by analytically marginalizing over power spectra amplitudes [117], or in procedure for characterizing the contribution of unresolved point sources by integrating over their positions [118], or in strong gravitational lensing analysis by marginalizing over the background source galaxy [119].

On the other hand, one key aspect of SBI algorithms in general is their ability to *directly estimate marginal posteriors*, instead of having to first estimate the joint posterior over all parameter space, and then marginalize out nuisance parameters. A clear example of this property can be found in Ref. [120], where both SBI and and a nested sampler are used to sample from an "eggbox" posterior with $D = 14$ dimensions with over 10^4 modes: the nested sample requires at least 10^7 samples, whereas SBI needs three order of magnitude less samples to directly estimate the 1-dimensional and 2-dimensional marginal posteriors for the 14 parameters [Figure 2 in 120].

A second illustrative example of this property can be found in Ref. [78]. Analyzing cosmic microwave background data, they used SBI to directly estimate marginal posteriors for the six cosmological parameters of interest, while directly marginalizing over the thirteen varying nuisance parameters present in the analysis. The SBI analysis required only 3×10^3 simulations, while to obtain the same results using MCMC with the Planck likelihood required 5×10^6 simulations (since the MCMC must sample all the nineteen parameters).

This, and many other comparative analysis of SBI and LBI methods, can be found in Figure 1.3. The figure is meant as a very simplified comparison of LBI and SBI algorithms in the space of number of required simulations (or computational time to run the analysis) versus number of model parameters. Each work addresses different physics applications, each with its own set of challenges and variations, employing diverse types of analysis to solve them. Therefore, each reported scatter point is unique and achieves the reported scaling thanks to different properties of the various SBI and LBI algorithms employed in the specific study. The main takeaway message from the plot is that, in general, SBI techniques are extremely scalable and simulation efficient with respect to LBI as a function of model parameters. One of the main reasons, except for the fast evaluation time of neural networks, is their possibility to directly estimate marginal probabilities. This effectively means that we can use SBI algorithms to break down large problems into smaller ones, while coherently accounting for the uncertainties coming from the rest of the parameter space (as further detailed in Section 2.2.2). In the next chapter, we will get deeper into the technical details of SBI and illustrate more reasons why these techniques are promising for pushing the boundaries of the curse of dimensionality compared to likelihood-based methods (Figure 1.4).

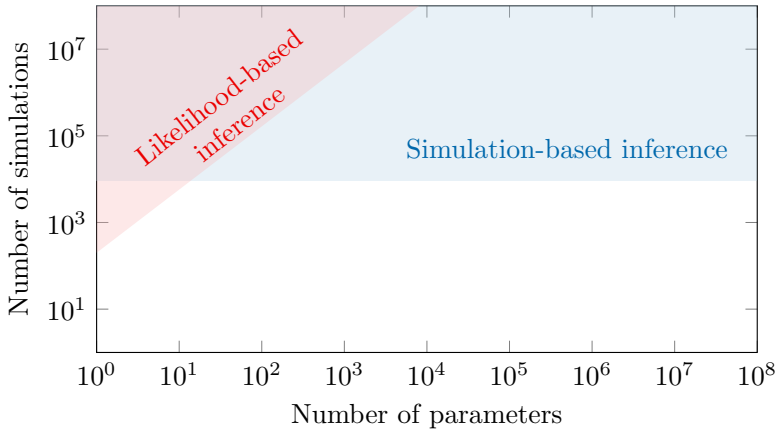

Figure 1.4: Simplified comparison of likelihood-based and simulation-based algorithms in the space of number of required simulations versus number of model parameters. In general, the simulation requirements of likelihood-based techniques grows significantly with the number of model parameters (curse of dimensionality). Instead, simulation-based inference techniques can, in principle, directly focus on estimating marginal posteriors for parameters of interest, independently of the total number of parameters. This reduces the need for parameter reduction techniques and enables the comparison of complex simulation results with complex data. The figure is adapted from Figure 4 in Ref. [121].

1.4 Outline

This thesis aims to contribute to the ongoing effort to transition towards simulation-based inference techniques in astrophysics and cosmology, emphasizing some of the tremendous opportunities that this transition brings. To this end, this thesis first proposes a general simulation-based ecosystem for astrophysical data analysis (Chapter 2). Then, it illustrate its capabilities through exemplary applications to three challenging astrophysical problems, as motivated in Section 1.1:

Chapters 3 and 4:	The analysis of strong gravitational lenses as a dark matter probe.
Chapter 5:	The reconstruction of cosmological initial conditions from late-time density fields.
Chapter 6:	The analysis of point-sources in sky-maps.

Overall, it aims to highlight the potential for fast, flexible, and testable simulation-based algorithms to facilitate scientific discovery in astrophysics and cosmology, at the dawn of their data-driven era, and forward.

2 Simulation-based inference

The purpose of this chapter is to complement the introduction by laying the foundations of the simulation-based statistical inference framework employed throughout the rest of the thesis. First, we provide an overview of traditional and neural network-based SBI implementations. Then, we focus on the specific algorithm that will be employed in the following chapters of this thesis, truncated marginal neural ratio estimation. We will then compare it with the other core SBI algorithms, highlighting advantages and pitfalls. Finally, we will outline this thesis specific contribution to the SBI field for astrophysical data analysis.

2.1 The landscape of SBI

In this section, we describe the developments and landscape of SBI algorithms. We begin in Section 2.1.1 by providing a brief overview of traditional implicit-likelihood methods for posterior inference, approximate bayesian computation [122] and kernel density estimation [123]. We then explain in Section 2.1.2 the different routes one can use to vastly accelerate and scale this process using neural networks [52, 124]. A comparison of the various neural SBI algorithms with a discussion about their drawbacks and advantages can be found later in Section 2.4.

2.1.1 Traditional SBI

The first implicit-likelihood ideas date back to the 1980s. Particularly relevant for the initial development of this paradigm was statistician Donald Rubin's discussion about the use of frequency calculations for the interpretation of Bayesian statements [125]. Importantly, in these lectures he encouraged statisticians to not settle for analytically tractable models only, but to instead consider computational methods to estimate the

posterior distribution of interest for a wider range of models. An outline of the most well-known traditional SBI algorithm, approximate Bayesian computation (ABC), can be already found in Rubin's work [125, Section 3.1].

Approximate bayesian computation

It is worth exploring in more details the ABC algorithm, since it will serve us as the classical analogue of the SBI technique mostly employed throughout this thesis (see Section 2.2). For more detailed reviews see Refs. [122, 126, 127]. ABC is a rejection sampling algorithm where, given a prior $\boldsymbol{\Theta} \sim p(\boldsymbol{\Theta})$ proposed samples $\boldsymbol{x} \sim p(\boldsymbol{x} \mid \boldsymbol{\Theta})$ from the forward model are compared to the target observed data $\boldsymbol{x}_0$ with a hand-crafted distance measure based on some low-dimensional summary statistics $s(\boldsymbol{x})$, as for example

$$d(\boldsymbol{x}, \boldsymbol{x}_0) = \|s(\boldsymbol{x}) - s(\boldsymbol{x}_0)\| \,. \tag{2.1}$$

Samples from the approximate posterior are drawn with rejection sampling using an acceptance tolerance ϵ such that samples satisfy $d(\boldsymbol{x}, \boldsymbol{x}_0) < \epsilon$. Hence the posterior

$$p_{\mathrm{ABC}}(\boldsymbol{\Theta} \mid \boldsymbol{x}) = \frac{\int_{d(\boldsymbol{x}_0, \boldsymbol{x}) < \epsilon} d\boldsymbol{x}\, p(\boldsymbol{\Theta} \mid \boldsymbol{x}) p(\boldsymbol{x})}{\int_{d(\boldsymbol{x}_0, \boldsymbol{x}) < \epsilon} d\boldsymbol{x}\, p(\boldsymbol{x})} \tag{2.2}$$

is guaranteed to converge to the true one for sufficiently informative summary statistics $s(\boldsymbol{x})$ and for $\epsilon \to 0$, $p_{\mathrm{ABC}}(\boldsymbol{\Theta} \mid \boldsymbol{x}) \xrightarrow{\epsilon \to 0} p(\boldsymbol{\Theta} \mid \boldsymbol{x})$. Whereas, when ϵ is non-zero, the approximate posterior is guaranteed to be broader than the true one, leading to conservative inference.

Interestingly, a physical implementation of ABC-rejection scheme for a single parameter and a single observation was already constructed by Francis Galton in the late 1800s [128, Figure 5]. The device is known today as the Galton board, and often used to demonstrate the central limit theorem [129]. Since Rubin's work, ABC has been widely adopted in astrophysics and cosmology, including applications to, e.g., galaxy demographics [130], galaxy–halo connection [131], intergalactic medium [132], exoplanetary systems [133], luminosity functions [134], stellar initial mass function [135], cosmology [136, 137], galaxy clustering [138], type Ia supernovae [139], gamma-ray sky [140], and cosmic rays [141]. The quick adoption of ABC from the astrophysical community has also been accompanied by the development of specialized softwares, which have played an important role in the dissemination of the technique such as `CosmoABC` [138], `abcpmc` [136], and `astroABC` [137].

Approximate frequentist computation

A second classical approach to SBI was also proposed in the 1980s by Diggle and Gratton [123]. The approach was dubbed "approximate frequentist computation" by the authors of Ref. [142] because of its similarities to ABC. Specifically, this method

is based on creating an approximate model for the likelihood by estimating the distribution of low-dimensional summary statistics from samples drawn from the simulator with histograms or *kernel density estimation*. The advantage over ABC is that it is amortized, meaning that after the initial computational cost for the simulation and density estimation phase, evaluating new data points becomes efficient. This property makes kernel density estimation-based inference particularly well suited for problems with many independent and identically distributed (i.i.d.) observations, a key reason for its widespread use in particle physics measurements [143]. For example, this approach was used for the discovery of the Higgs boson in a frequentist paradigm [142].

Moving SBI forward

Despite their success and continuous progresses, there are some explicit disadvantages of these classical SBI methods (of which the community is aware of [see e.g. 24]), that makes them unsuitable for high-dimensional and complex data. First, they both rely on low-dimensional summary statistics $s(\boldsymbol{x})$ to compare simulations to the data, which may not retain all the information available in the data. Second, they suffer from the curse of dimensionality, with a required simulation budget that increases with the dimensionality of the parameter space [e.g. 115]. In particular, the acceptance rate of ABC's rejection algorithm vanishes exponentially as the dimensionality of the parameter space increases, thus significantly more simulation budget is needed to reach convergence in high dimensions for $\epsilon \to 0$.

These challenges are tackled in modern SBI algorithms thanks to advances in deep learning [53] and automatic differentiation [54]. In particular, first, the development of neural network's architectures tailored to various data structures has been fundamental for the processing of significantly more complex data [53]. This allows for the optimization of learned data features w.r.t. some custom loss, instead of relying on hand-crafted summary statistics. Second, neural network-based algorithms are being actively developed to estimate probability density distributions in high dimensions, overcoming the curse of dimensionality [e.g. 144–146]. In the next section, we will discuss the main neural network-based SBI algorithms.

2.1.2 Neural SBI

Moving beyond the classical SBI approaches, in recent years, a number of new neural network-based SBI techniques have been proposed. We refer the reader to Ref. [52] for a general review, to Ref. [124] for benchmarks of the various algorithms, and to Ref. [147] for an exhaustive comparisons of all SBI methods currently available as well as discussions about their challenges and pitfalls in the context of astrophysics and

cosmology.[1] The widespread adoption of these methods in astrophysics and cosmology has also been driven by the development of specialized tools, such as, e.g., `pydelfi` [55], `sbi` [56], `swyft` [57], and `lampe` [58].

In general, Bayes' theorem (Equation (1.1)) hints at a few different approaches for the implementation of neural SBI algorithms: neural posterior estimation (NPE) which employs density estimation techniques to directly estimate the posterior $p(\boldsymbol{\Theta} \mid \boldsymbol{x})$ [146, 148]; neural likelihood estimation (NLE) which instead uses density estimation to learn an approximation to the likelihood $p(\boldsymbol{x} \mid \boldsymbol{\Theta})$ [149, 150]; and neural ratio estimation (NRE) which uses classifiers to approximate the likelihood-to-evidence or posterior-to-prior ratio $\frac{p(\boldsymbol{x}|\boldsymbol{\Theta})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{\Theta}|\boldsymbol{x})}{p(\boldsymbol{\Theta})}$ [120, 151–154]. In the following paragraphs, we will describe their fundamental properties.

Neural posterior estimation

NPE is the most straightforward method [146, 148]. It introduces a density estimator $q_\Phi^{\mathrm{NPE}}(\boldsymbol{\Theta} \mid \boldsymbol{x})$, parametrized through the weights Φ, which is trained to approximate the posterior for parameters $\boldsymbol{\Theta}$ given data $\boldsymbol{x}$,

$$q_\Phi^{\mathrm{NPE}}(\boldsymbol{\Theta} \mid \boldsymbol{x}) \approx p(\boldsymbol{\Theta} \mid \boldsymbol{x}) \,. \tag{2.3}$$

Specifically, we want to reduce the dissimilarity between the density estimator $q_\Phi^{\mathrm{NPE}}(\boldsymbol{\Theta} \mid \boldsymbol{x})$ and the true posterior $p(\boldsymbol{\Theta} \mid \boldsymbol{x})$ by minimizing w.r.t. the weights Φ the *forward* Kullback-Leibler (KL) divergence [155],

$$D_{\mathrm{KL}}(p \| q_\Phi^{\mathrm{NPE}}) = \int \mathrm{d}\boldsymbol{\Theta}\, p\left(\boldsymbol{\Theta} \mid \boldsymbol{x}\right) \ln \left(\frac{p\left(\boldsymbol{\Theta} \mid \boldsymbol{x}\right)}{q_\Phi^{\mathrm{NPE}}\left(\boldsymbol{\Theta} \mid \boldsymbol{x}\right)} \right) \tag{2.4}$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{\Theta} \sim p(\boldsymbol{\Theta}|\boldsymbol{x})} \left[\ln \frac{p\left(\boldsymbol{\Theta} \mid \boldsymbol{x}\right)}{q_\Phi^{\mathrm{NPE}}\left(\boldsymbol{\Theta} \mid \boldsymbol{x}\right)} \right] \tag{2.5}$$

As the expectation value is taken over samples from the true posterior distribution, we cannot evaluate the above quantity. To overcome this limitation, one averages over simulation data $\boldsymbol{x} \sim p(\boldsymbol{x})$. The loss is now the expected forward KL divergence, where we have exploited the fact that $p(\boldsymbol{\Theta} \mid \boldsymbol{x})p(\boldsymbol{x}) = p(\boldsymbol{x}, \boldsymbol{\Theta})$, which we know how to sample:

$$\mathop{\mathbb{E}}_{\boldsymbol{x} \sim p(\boldsymbol{x})} \left[D_{\mathrm{KL}}(p \| q_\Phi^{\mathrm{NPE}}) \right] = \mathop{\mathbb{E}}_{\boldsymbol{x}, \boldsymbol{\Theta} \sim p(\boldsymbol{x}, \boldsymbol{\Theta})} \left[\ln \frac{p(\boldsymbol{\Theta} \mid \boldsymbol{x})}{q_\Phi^{\mathrm{NPE}}(\boldsymbol{\Theta} \mid \boldsymbol{x})} \right] \tag{2.6}$$

$$= \underbrace{- \mathop{\mathbb{E}}_{\boldsymbol{x}, \boldsymbol{\Theta} \sim p(\boldsymbol{x}, \boldsymbol{\Theta})} \ln q_\Phi^{\mathrm{NPE}}(\boldsymbol{\Theta} \mid \boldsymbol{x})}_{\text{expected entropy}} + \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{x}, \boldsymbol{\Theta} \sim p(\boldsymbol{x}, \boldsymbol{\Theta})} \ln p(\boldsymbol{\Theta} \mid \boldsymbol{x})}_{\text{constant}} \,. \tag{2.7}$$

[1] An extensive list of works that use neural SBI in astrophysics, cosmology, and high energy physics can be found here: https://github.com/smsharma/awesome-neural-sbi.

NPE loss function is then given by the negative log-probability

$$\mathcal{L}^{\mathrm{NPE}} = - \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{\Theta}\sim p(\boldsymbol{x},\boldsymbol{\Theta})} \ln q_\Phi^{\mathrm{NPE}}(\boldsymbol{\Theta} \mid \boldsymbol{x}) \,. \tag{2.8}$$

Application of this loss function requires that the density estimator is normalized to one, $\int \mathrm{d}\boldsymbol{\Theta}\; q_\Phi^{\mathrm{NPE}}(\boldsymbol{\Theta} \mid \boldsymbol{x}) = 1$, which can be guaranteed by using specific network architectures to parametrize q_Φ^{NPE}. The most popular ones with this property are normalizing flows, which transform tractable density functions into complex ones through a sequence of invertible operations with tractable Jacobians [145, 156]. The NPE algorithm has the advantage that it allows both for evaluation and sampling of the estimated posterior from the normalizing flow. However, the pitfall is that it requires an analytic form of the proposal prior from which training data was sampled, which may not be always accessible (e.g. from cosmological simulations).

NPE has been widely used in astrophysics, e.g., for strong gravitational lensing analysis [67–69], gravitational waves parameter estimation [59–61], gravitational wave background reconstruction [62], cosmological parameter inference [72], X-ray spectral fitting [83], exoplanetary atmospheric retrieval [84], galaxy spectra inference [85, 86], galactic center γ-ray excess characterization [87, 88], among other applications.

Comparison with variational inference It is interesting to compare NPE with its direct likelihood-based counterpart, variational inference (VI), already mentioned in Section 1.3 (for a review see Ref. [105]). VI allows the approximation of extremely high-dimensional Bayesian posteriors with simple proposal distributions by solving an optimization problem. This is achieved by using a proposal distribution, the so-called variational posterior, $q_\Phi^{\mathrm{VI}}(\boldsymbol{\Theta} \mid \boldsymbol{x}_0)$, to approximate the true posterior $p\,(\boldsymbol{\Theta} \mid \boldsymbol{x}_0)$ and optimizing its parameters Φ using gradient ascent [51, 157, 158]. VI's optimization objective is the *reverse* KL divergence

$$D_{\mathrm{KL}}(q_\Phi^{\mathrm{VI}} \| p) = \int \mathrm{d}\boldsymbol{\Theta}\; q_\Phi^{\mathrm{VI}}\,(\boldsymbol{\Theta} \mid \boldsymbol{x}_0) \ln \left(\frac{q_\Phi^{\mathrm{VI}}\,(\boldsymbol{\Theta} \mid \boldsymbol{x}_0)}{p\,(\boldsymbol{\Theta} \mid \boldsymbol{x}_0)} \right) \tag{2.9}$$

$$= \mathop{\mathbb{E}}_{\boldsymbol{\Theta}\sim q_\Phi^{\mathrm{VI}}(\boldsymbol{\Theta}|\boldsymbol{x}_0)} \left[\ln \frac{q_\Phi^{\mathrm{VI}}\,(\boldsymbol{\Theta} \mid \boldsymbol{x}_0)}{p\,(\boldsymbol{\Theta} \mid \boldsymbol{x}_o)} \right] \tag{2.10}$$

Similarly to the trick used for NPE, in order to avoid evaluating the inaccessible posterior $p\,(\boldsymbol{\Theta} \mid \boldsymbol{x}_o)$, one exploits Bayes' theorem and substitute $p(\boldsymbol{\Theta}|\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})/p(\boldsymbol{x})$, obtaining:

$$D_{\mathrm{KL}}(q_\Phi^{\mathrm{VI}} \| p) = \mathop{\mathbb{E}}_{\boldsymbol{\Theta}\sim q_\Phi^{\mathrm{VI}}(\boldsymbol{\Theta}|\boldsymbol{x}_o)} \left[\ln \frac{q_\Phi^{\mathrm{VI}}(\boldsymbol{\Theta} \mid \boldsymbol{x}_o)}{p(\boldsymbol{\Theta} \mid \boldsymbol{x}_o)} \right] \tag{2.11}$$

$$= \underbrace{\mathop{\mathbb{E}}_{\boldsymbol{\Theta}\sim q_\Phi^{\mathrm{VI}}(\boldsymbol{\Theta}|\boldsymbol{x}_o)} \left[\ln \frac{q_\Phi^{\mathrm{VI}}(\boldsymbol{\Theta} \mid \boldsymbol{x}_o)}{p(\boldsymbol{x}_0 \mid \boldsymbol{\Theta}) p(\boldsymbol{\Theta})} \right]}_{\equiv -\mathrm{ELBO}} + \underbrace{\ln p(\boldsymbol{x}_0)}_{\text{constant}} \,, \tag{2.12}$$

where we have introduced the evidence lower bound (ELBO). Minimizing the reverse KL divergence is thus equivalent to maximizing the ELBO, and the VI loss is

$$\mathcal{L}^{\text{VI}} = -\text{ELBO} \equiv \mathop{\mathbb{E}}_{\boldsymbol{\Theta}\sim q_{\Phi}^{\text{VI}}(\boldsymbol{\Theta}|\boldsymbol{x}_0)} \left[\ln p\left(\boldsymbol{x}_0, \boldsymbol{\Theta}\right) - \ln q_{\Phi}^{\text{VI}}\left(\boldsymbol{x}_0 \mid \boldsymbol{\Theta}\right)\right] . \tag{2.13}$$

We can now comment upon the main differences between NPE and VI. First, in NPE parameters are drawn from the model prior $p(\boldsymbol{\Theta})$ (see Equation (2.8)), whereas in VI they are drawn from the variational posterior $q_{\Phi}^{\text{VI}}\left(\boldsymbol{\Theta} \mid \boldsymbol{x}_0\right)$ (see Equation (2.13)). As a result, in the latter case, samples increasingly focus on regions with high data likelihood $p(\boldsymbol{x}_o \mid \boldsymbol{\Theta})$, whereas in NPE there is no automatic focusing on regions with high data likelihood. Second, for VI the posterior $q_{\Phi}^{\text{VI}}\left(\boldsymbol{\Theta} \mid \boldsymbol{x}_0\right)$ must cover all parameters that the likelihood model is conditioned on, and omitting correlations (e.g. mean-field approximation) gives a lower bound on parameter uncertainties. On the other hand, in NPE, if we substitute $q_{\Phi}^{\text{NPE}}(\boldsymbol{\Theta} \mid \boldsymbol{x})$ with a low dimensional marginal, e.g. $q_{\Phi}^{\text{NPE}}(\Theta_1 \mid \boldsymbol{x})$, we automatically get marginal posterior estimates, without performing integrals explicitly. Hence, omitting correlations generates correctly marginalized posteriors [159]. The ability to directly perform marginal inference is shared among all other neural SBI algorithms, and it is one of the main differences with likelihood-based algorithms.

Neural likelihood estimation

Another choice for neural SBI is to perform NLE [149, 150], where one trains an estimator for the likelihood probability density of the data $\boldsymbol{x}$ given some model parameters $\boldsymbol{\Theta}$,

$$q_{\Phi}^{\text{NLE}}(\boldsymbol{x} \mid \boldsymbol{\Theta}) \approx p(\boldsymbol{x} \mid \boldsymbol{\Theta}) . \tag{2.14}$$

Following a similar reasoning as the one employed to derive NPE loss function, NLE loss function is given by the negative log-likelihood

$$\mathcal{L}^{\text{NLE}} = - \mathop{\mathbb{E}}_{\boldsymbol{x},\boldsymbol{\Theta}\sim p(\boldsymbol{x},\boldsymbol{\Theta})} \ln q_{\Phi}^{\text{NLE}}(\boldsymbol{x} \mid \boldsymbol{\Theta}) . \tag{2.15}$$

Again, specialized network architectures as normalizing flows are typically employed to guarantee that $q_{\Phi}^{\text{NLE}}(\boldsymbol{x} \mid \boldsymbol{\Theta})$ is a properly normalized density function.

As a general guideline, the choice between NPE or NLE for different physical problems should consider the dimensionality of the data and parameter space, as well as the complexity of learning likelihood or posterior distribution's shape. Indeed, neural networks are easier to train when using high-dimensional input to yield low-dimensional output than vice versa. Thus, if the data is high-dimensional (e.g., images), then NPE may be a more suitable choice than NLE. Conversely, for high-dimensional posterior inference (i.e., when the dimension of the parameter space is

large), NLE often outperforms NPE, especially when there are strong parameter degeneracies.

NLE has also been widely used in astrophysics, for gravitational waves parameter estimation [63] and cosmological parameter inference [55, 73–77], among other applications.

Neural ratio estimation

Bayes' theorem hints at a last quantity that can be estimated, the likelihood-to-evidence or posterior-to-prior ratio

$$r(\boldsymbol{\Theta}; \boldsymbol{x}) \approx \frac{p(\boldsymbol{x} \mid \boldsymbol{\Theta})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{\Theta} \mid \boldsymbol{x})}{p(\boldsymbol{\Theta})} \,. \tag{2.16}$$

This quantity is approximated using binary classification via a supervised learning task. In this case, the ratio does not need to be normalized, hence one is free to use any neural network architecture.

We will explore in more details the NRE algorithm in a dedicated section, Section 2.2.1, as in this thesis we will focus on this last method and its extensions. Moreover, a detailed comparison between NRE and its extension with the other neural SBI algorithms is presented in Section 2.4.

NRE has also been widely used in astrophysics and cosmology, for gravitational waves parameter estimation [64, 65], gravitational wave background reconstruction [66], strong gravitational lensing analysis [1, 5, 70, 71], cosmology [3, 78, 79], stellar streams analysis [89, 90], 21-cm power spectra [80], type Ia supernovae [81, 82], point source detection [4], millisecond pulsars [91], astrometry [92], among other applications.

Active learning

There are a number of additional features that can be used on top of these core SBI algorithms, either preceding or following the main inference step. Here, we further refine the taxonomy of SBI algorithms based on whether they use some form of active learning to guide the simulator towards parts of the parameter space that are most relevant for a specific target observation $\boldsymbol{x}_0$. In particular, *sequential* SBI approaches adaptively choose informative simulations by using sequentially refined proposal distributions for the model parameters. They have been reported to outperform and be more simulation-efficient with respect to non-sequential ones across a number of different benchmark tasks [124] and, e.g., in cosmology [78], in gravitational waves astrophysics [64], and in strong lensing analysis [69]. Particularly impressive is the recent result form Ref. [69], where they perform a sequential NPE analysis of strong gravitational lenses, that would have taken 2000 times longer without the use of active learning [Figure 5 in 69].

The core intuition behind sequential SBI algorithms is the following: given a single observation of interest, $\boldsymbol{x}_0$, sampling parameters from the entire prior space to generate training data may not be efficient, since it leads to training data $(\boldsymbol{x}, \boldsymbol{\Theta})$ that has significant variance compared to the target observation $\boldsymbol{x}_0$. Therefore, for a fixed simulation budget, the training samples contain only limited information about the posterior $p(\boldsymbol{\Theta} \mid \boldsymbol{x}_0)$. The alternative proposed by sequential algorithms, in order to increase simulation efficiency, is to draw parameters from an adaptive proposal distribution $p^{(R)}(\boldsymbol{\Theta})$, resulting in training data that matches the observation of interest more closely in each sequential round R. More details regarding various possible implementations of sequential inference are given in Section 2.2.3.

In this section we have introduced the core neural SBI algorithms. After describing in details the algorithm mostly employed in this thesis (Section 2.2), we will compare it to the other ones in Section 2.4.

2.2 Truncated Marginal Neural Ratio Estimation

We will now focus our attention on the main SBI technique employed in this thesis, truncated marginal neural ratio estimation (TMNRE). We provide a pedagogical introduction, and show that it forms the core of a fast, accurate and precise approach to solve general parameter inference problems in high-dimensional settings [120, 160, 161]. The approach builds on three key simple ingredients. First, neural ratio estimation (NRE). Second, focus on marginal inference (M). Third, active learning through prior truncation (T). We provide a technical description of them in the following sections, emphasizing how they compose well together.

2.2.1 Neural Ratio Estimation: "classification is all you need"

Ratio estimation rephrases Bayesian posterior inference as a binary classification problem. Given an implicitly defined model $\boldsymbol{x}, \boldsymbol{\Theta} \sim p(\boldsymbol{x}, \boldsymbol{\Theta}) = p(\boldsymbol{x} \mid \boldsymbol{\Theta})p(\boldsymbol{\Theta})$, the idea behind ratio estimation is to use a binary classifier to distinguish between data $\boldsymbol{x}$ and parameter $\boldsymbol{\Theta}$ pairs drawn from two classes labeled by the binary variable C:

$$p(\boldsymbol{x}, \boldsymbol{\Theta} \mid C = 1) = p(\boldsymbol{x}, \boldsymbol{\Theta}) \tag{2.17}$$

$$p(\boldsymbol{x}, \boldsymbol{\Theta} \mid C = 0) = p(\boldsymbol{x})p(\boldsymbol{\Theta})\,. \tag{2.18}$$

Concretely, these two distributions correspond respectively to drawing data and parameters jointly from the simulator, $\boldsymbol{x}, \boldsymbol{\Theta} \sim p(\boldsymbol{x}, \boldsymbol{\Theta})$, or to drawing data and parameters marginally, $\boldsymbol{x}, \boldsymbol{\Theta} \sim p(\boldsymbol{x})p(\boldsymbol{\Theta})$, by sampling an unrelated set of parameters from

the prior versus data from the simulator (this can be simply obtained by scrambling the joint pairs).[2]

Sampling $C = 0$ and $C = 1$ with equal probability, the decision function for the Bayes-optimal classifier [162] (i.e. the one that minimizes the Bayesian risk of missclassification) is:

$$p(C = 1 \mid \boldsymbol{x}, \boldsymbol{\Theta}) = \frac{p(\boldsymbol{x}, \boldsymbol{\Theta})}{p(\boldsymbol{x}, \boldsymbol{\Theta}) + p(\boldsymbol{x})p(\boldsymbol{\Theta})} \equiv \sigma[\log r(\boldsymbol{\Theta}; \boldsymbol{x})] \,, \tag{2.19}$$

where we introduced the sigmoid function $\sigma(y) \equiv 1/(1 + e^{-y})$. Importantly, the equivalence in Equation (2.19), often called the "likelihood ratio trick" [e.g. 70, 151], shows how the Bayes-optimal classifier is related to the posterior-to-prior, likelihood-to-evidence, or joint-to-marginal distribution ratio:

$$r(\boldsymbol{\Theta}; \boldsymbol{x}) \equiv \frac{p(\boldsymbol{\Theta} \mid \boldsymbol{x})}{p(\boldsymbol{\Theta})} = \frac{p(\boldsymbol{x} \mid \boldsymbol{\Theta})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}, \boldsymbol{\Theta})}{p(\boldsymbol{x})p(\boldsymbol{\Theta})} \,. \tag{2.20}$$

Thus, one can gain access to the posterior-to-prior ratio by training a classifier of joint versus marginal pairs, which are easily obtainable with a forward simulator, and subsequently use it for inference. In practice, as the first equality in Equation (2.20) suggests, if the prior is tractable, $r(\boldsymbol{\Theta}; \boldsymbol{x})$ gives direct access to the posterior density: $p(\boldsymbol{\Theta} \mid \boldsymbol{x}) = r(\boldsymbol{\Theta}; \boldsymbol{x})p(\boldsymbol{\Theta})$. Alternatively, one can use $r(\boldsymbol{\Theta}; \boldsymbol{x})$ to weight prior samples, enabling posterior sampling even when the prior cannot be expressed in closed-form.

Often, both the data and the parameters are typically high-dimensional objects (e.g. strong lensing images and the parameters of a population of dark matter substructures, or galaxy clustering measures and cosmological initial conditions parameters), and therefore neural network-based classifiers that are able to process this complex data and be efficiently optimized are a necessary choice. As proposed by Ref. [153], in NRE the classifier in Equation (2.19) is a neural network (NN), $d_{\boldsymbol{\Phi}}(\boldsymbol{x}, \boldsymbol{\Theta})$, that takes as input a parameters-data pair and produces an estimate of the ratio $\hat{r}(\boldsymbol{\Theta}; \boldsymbol{x})$.[3] The network parameters $\boldsymbol{\Phi}$ are optimized via stochastic gradient descent [51, 157, 158] to minimize the binary cross-entropy (BCE) [163]:

$$\begin{aligned} \mathcal{L}[d_{\boldsymbol{\Phi}}(\boldsymbol{x}, \boldsymbol{\Theta})] = -\int \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{\Theta} \, \{ & p(\boldsymbol{x}, \boldsymbol{\Theta}) \log d_{\boldsymbol{\Phi}}(\boldsymbol{x}, \boldsymbol{\Theta})] \\ & + p(\boldsymbol{x})p(\boldsymbol{\Theta}) \log \left[1 - d_{\boldsymbol{\Phi}}(\boldsymbol{x}, \boldsymbol{\Theta})\right] \} \,. \end{aligned} \tag{2.21}$$

Therefore, by training a neural network $d_{\boldsymbol{\Phi}}(\boldsymbol{x}, \boldsymbol{\Theta})$ to estimate $\hat{r}(\boldsymbol{\Theta}; \boldsymbol{x})$ via this supervised classification task, we obtain an estimate of the posterior through $\hat{p}(\boldsymbol{\Theta} \mid \boldsymbol{x}) = \hat{r}(\boldsymbol{\Theta}; \boldsymbol{x})p(\boldsymbol{\Theta})$.

[2] Throughout this thesis we will refer to joint samples as *positive* training examples, and to marginal samples as *negative* training examples.

[3] Throughout this chapter we will use the notation $\hat{\square}$ to indicate quantities estimated via NN.

2.2.2 Marginalization: focus on the essentials

Joint posteriors are commonly a key component of a scientific workflow when analyzing real-world data, since access to joint posterior samples exhaustively solves a given Bayesian parameter inference. However, when the number of parameters becomes large (e.g., hundreds for a cosmic shear joint analysis of three surveys [164], thousands for strong gravitational lensing substructures [1, 5] or background source reconstruction [41], or even millions for cosmological initial conditions [3, 44]), the computational overhead of generating joint samples can be enormous. Furthermore, the full joint posterior $p(\boldsymbol{\Theta} \mid \boldsymbol{x})$ is usually only an intermediate step, and not the goal by itself: in many cases, scientific insight is based on a low-dimensional marginalization of the overly informative joint posterior over nuisance parameters.

Let us consider a model with a full joint distribution

$$p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\eta}) = p(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{\eta}) p(\boldsymbol{\theta}, \boldsymbol{\eta}) \,, \tag{2.22}$$

where we have split the set of all model parameters $\boldsymbol{\Theta} = \{\boldsymbol{\theta}, \boldsymbol{\eta}\} \in \mathbb{R}^D$ into parameters of interest $\boldsymbol{\theta} \in \mathbb{R}^d$ (which we want to infer) and nuisance parameters $\boldsymbol{\eta} \in \mathbb{R}^{D-d}$ (which we want to marginalize over). Formally, the marginal posterior is obtained by integrating the joint posterior over all nuisance parameters

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}) = \int \mathrm{d}\boldsymbol{\eta} \; p(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \boldsymbol{x}) \,. \tag{2.23}$$

Given the general NRE setup, the extension to estimating marginal posteriors is straightforward: parameters to be marginalized over must be *sampled*, but *not presented* to the ratio estimator [120]. Hence, marginalization over nuisance variables is done implicitly, since the data will incorporate the variance from the nuisance parameters, but the inference procedure estimates only the marginal likelihood-to-evidence ratio.

In more detail, if $\boldsymbol{\eta}$ is not passed to the ratio estimator, the loss function becomes

$$\begin{aligned} \mathcal{L}[d_{\boldsymbol{\Phi}}(\boldsymbol{x}, \boldsymbol{\theta})] = -\int \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{\theta} \mathrm{d}\boldsymbol{\eta} \, \{ & p(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\eta}) \log d_{\boldsymbol{\Phi}}(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\eta}) \\ & + p(\boldsymbol{x}) p(\boldsymbol{\theta}, \boldsymbol{\eta}) \log \left[1 - d_{\boldsymbol{\Phi}}(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\eta})\right] \} \end{aligned} \tag{2.24}$$

$$\begin{aligned} = -\int \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{\theta} \, \{ & p(\boldsymbol{x}, \boldsymbol{\theta}) \log d_{\boldsymbol{\Phi}}(\boldsymbol{x}, \boldsymbol{\theta}) \\ & + p(\boldsymbol{x}) p(\boldsymbol{\theta}) \log \left[1 - d_{\boldsymbol{\Phi}}(\boldsymbol{x}, \boldsymbol{\theta})\right] \} \,, \end{aligned} \tag{2.25}$$

where we have integrated over $\boldsymbol{\eta}$ to obtain the second equality, proving our statement. As a result, the binary classifier trained on $(\boldsymbol{x}, \boldsymbol{\theta})$ pairs effectively learns an estimate of the marginal likelihood-to-evidence ratios

$$\hat{r}(\boldsymbol{x}; \boldsymbol{\theta}) \equiv \frac{p(\boldsymbol{\theta} \mid \boldsymbol{x})}{p(\boldsymbol{\theta})} = \int \mathrm{d}\boldsymbol{\eta} \, \frac{p(\boldsymbol{\theta}, \boldsymbol{\eta} \mid \boldsymbol{x})}{p(\boldsymbol{\theta})} p(\boldsymbol{\eta}) \,. \tag{2.26}$$

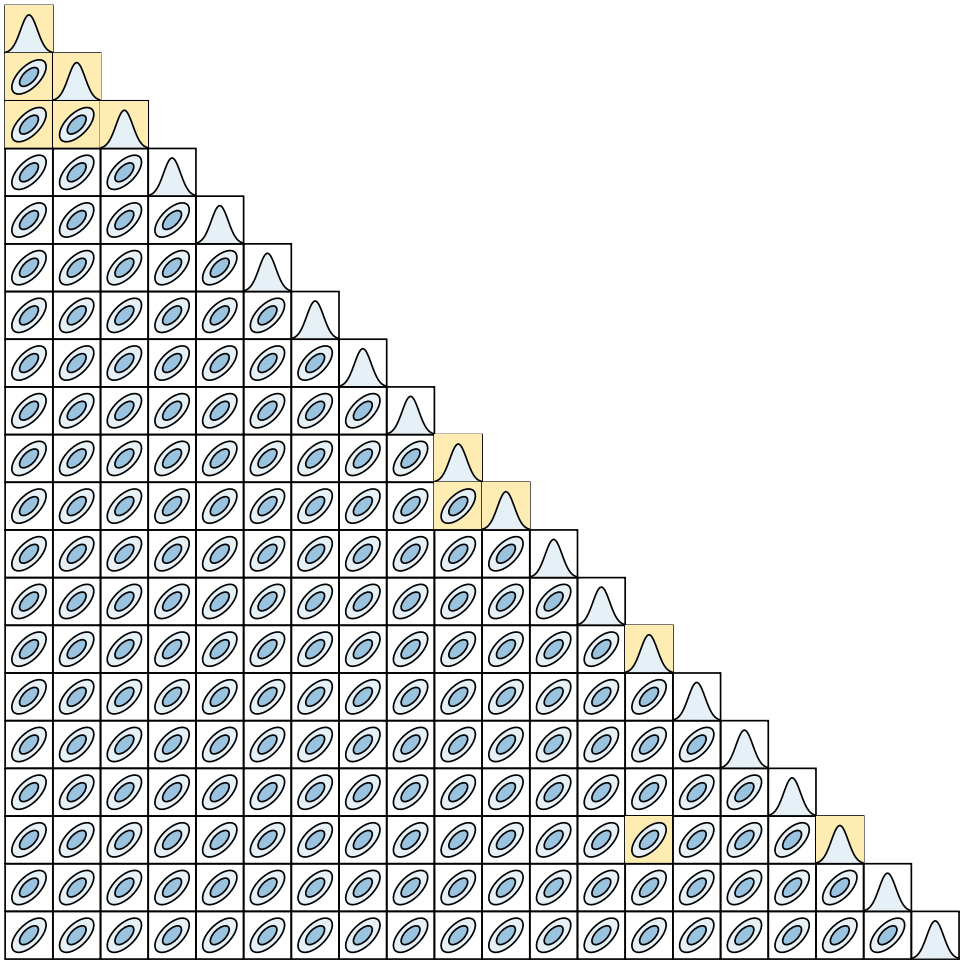

Figure 2.1: In simulation-based setting it is possible to directly estimate the marginal posterior of interest, instead of solving the problem for the full joint posterior at once, which can be computationally infeasible for complex models. Here we highlight some of the possible types of marginals that can be directly inferred in SBI for this mock high-dimensional parameter space.

We can then use the marginal ratio $\hat{r}(\boldsymbol{x}, \boldsymbol{\theta})$ to evaluate the marginal posterior for the parameters of interest directly or obtain samples otherwise, with significantly less computational cost. The possibility to directly access marginal posteriors can be easily obtained for NLE and NPE density estimators as well [55, 165].

As touched upon in Section 1.3, this way of approaching the problem determines the scalability of SBI algorithms with parameter space dimensionality. One can then focus on improving the model realism (i.e. complexifying the simulator) without fundamentally altering the inference process (i.e. same ratio estimator training), since there is no need to estimate a full-joint posterior. Effectively, this means that we can use SBI algorithms to break down large problems into smaller ones, while coherently accounting for the uncertainties coming from the rest of the parameter space. An exemplification of this process is illustrated in Figure 2.1, where instead of solving the inference problem for the whole parameter space, one can focus on lower-dimensional projections of the correlations.

It is important to note that, when making the simulator more complex by adding parameters, if all of them contribute equally to the data variance, the implicit data distribution will become noise-dominated. Thus, when referring to scaling to arbitrary number of variables, the data variance is implicitly kept fixed. This limit remains a challenge for likelihood-based methods, but is tractable in the simulation-based frameworks.

2.2.3 Truncation: zoom-in for high precision

The ratio estimators discussed so far are fully *amortized*: that is, they attempt to learn $r(\boldsymbol{\theta}; \boldsymbol{x})$ over the whole range of the prior $p(\boldsymbol{\theta})$. In principle, it is useful to be able to analyze any possible observation with the same network. In practice, when the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x}_0)$ for a particular observation $\boldsymbol{x}_0$ is much narrower than the prior, training an accurate ratio estimator, and general density estimators, requires a massive amount of training data. Hence, for a given limited simulation budget and network bandwidth, amortized inference typically comes at the expense of reduced posterior precision. In order to fully exploit available information in the data with limited computational resources, we instead focus on the problem of *targeted learning* of the posterior for a specific observation of interest $\boldsymbol{x}_0$ through *sequential inference*. In a nutshell, sequential inference approaches adaptively choose informative simulations by using sequentially refined proposal distributions for the model parameters, as briefly mentioned in Section 2.1. In this way, the relevant parameter space is sampled more densely and the network can learn better in that region.

The classification of sequential SBI algorithms can be sifted based on how they acquire new, informative simulations. In particular, the sequential techniques adopted

in Refs. [146, 148, 149, 153, 166, 167] draw new simulations for the next round from the *approximate posteriors* learned in each round. However, this approach suffers from two limitations. First, several frequently used diagnostic tools for SBI depend on performing inference across multiple observations (e.g. expected coverage tests, see Section 2.3). In this setting, to perform these tests, one would have to generate new simulations and network retraining for each observation, which is often prohibitively expensive. Second, marginal posterior estimation is in general affected by the proposal distribution, since one implicitly integrates over it. As a result, this approach is unsuitable for learning multiple marginal posteriors simultaneously over a number of sequential rounds, since sampling from the marginal for a particular parameter may hinder learning the marginals for other parameters (for a possible workaround see Ref. [55]).

To overcome the limitations of this sequential scheme, Ref. [160] proposed a hard-likelihood *prior truncation* scheme, applicable to NRE, that composes well with marginalisation and is locally amortised.[4] This prior truncation scheme iteratively discards in rounds R low likelihood(-to-evidence) regions, where the current approximate likelihood-to-evidence ratio evaluated for the target observation is below a user defined threshold ϵ. Concretely, this means keeping the region of parameter space defined by

$$\Gamma_{\boldsymbol{\theta}}^{(R)} = \{\boldsymbol{\theta} \in \mathbb{R}^d : \hat{r}^{(R)}(\boldsymbol{\theta}; \boldsymbol{x}) > \epsilon\} \,, \tag{2.27}$$

and discarding its complement.[5] Such truncated priors constrained to the parameter region $\Gamma_{\boldsymbol{\theta}}^{(R)}$ can be defined as:

$$p_\Gamma(\boldsymbol{\theta}) \equiv \frac{1}{V} \mathbb{I}_\Gamma(\boldsymbol{\theta}) p(\boldsymbol{\theta}) \,. \tag{2.28}$$

Here, $\mathbb{I}_\Gamma(\boldsymbol{\theta})$ is an indicator function, which is one for $\boldsymbol{\theta} \in \Gamma$ and zero otherwise,

$$\mathbb{I}_\Gamma(\boldsymbol{\theta}) \equiv \begin{cases} 1 & \text{for} \quad \boldsymbol{\theta} \in \Gamma \\ 0 & \text{otherwise} \end{cases} \tag{2.29}$$

Furthermore, $V \equiv \int d\boldsymbol{\theta}\, \mathbb{I}_\Gamma(\boldsymbol{\theta})\, p(\boldsymbol{\theta})$ is a normalizing constant that can be interpreted as the mass of the truncated prior.

The prior truncation scheme thus defines a series of nested indicator functions $\mathbb{I}_{\Gamma^{(R)}}$ whose regions have the property $\Gamma^{(0)} \supset \Gamma^{(1)} \supset \Gamma^{(2)} \supset \cdots \supset \Gamma^{(S)}$, where $\Gamma^{(0)}$ defines the support of the full initial prior, and $\Gamma^{(S)}$ is the final stable truncated prior. We illustrate this truncation procedure for a simple 1-dimensional scenario in Figure 2.2.

[4] With *locally amortised* inference we refer to an inference that can be repeated several times, without retraining, with distinct observations that live in the support of the truncated prior.

[5] Similar truncated proposals have also been introduced in Ref. [168] in the context of NPE, where the condition is instead on the current estimated posterior, e.g. $\hat{p}^{(R)}(\boldsymbol{\theta} \mid \boldsymbol{x}) > \epsilon$.

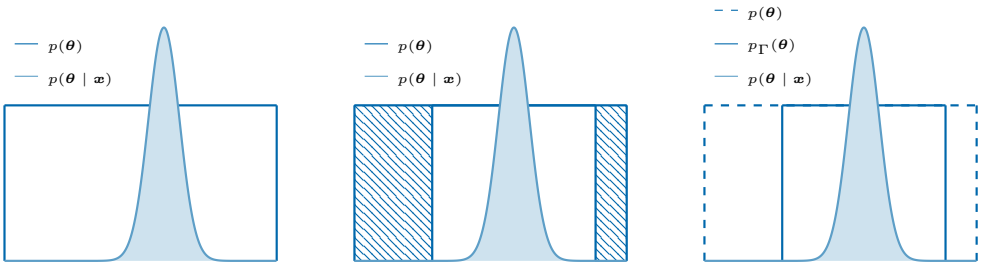

Figure 2.2: Illustration of sequential inference for a 1-dimensional posterior. First, we learn an approximation to the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ from the full initial prior $p(\boldsymbol{\theta})$. Then, we restrict the prior in the region where the parameter $\boldsymbol{\theta}$ is likely to have generated $\boldsymbol{x}$. In the next inference round, the truncated prior $p_\Gamma(\boldsymbol{\theta})$ will be used to generate targeted simulations, sampling the parameter space more densely in the region of interest, so that the network can learn better in that region.

Importantly, since this sequential scheme does not modify the shape of the prior proposal distribution, but only restricts its support, the inference is still locally amortised in the constrained proposal distribution region, thus allowing for the possibility of tests that rely on performing inference across multiple observations (in the Γ support). Moreover, it is also possible to use the same training data generated for a round to efficiently train arbitrary marginal posteriors for any set of model parameters.

Training targeted ratio estimators is achieved by replacing the full prior with a *truncated prior* $p_\Gamma^{(R)}(\boldsymbol{\theta})$, where the parameters are restricted to a region Γ where they are likely to have generated $\boldsymbol{x}_0$. Since parameters from the complement of Γ are unlikely to have generated $\boldsymbol{x}_0$, training a ratio estimator with data generated from the truncated prior as opposed to the full prior has little impact on the posterior learned by our ratio estimators. Indeed, only those regions where $p(\boldsymbol{x}_0 \mid \boldsymbol{\theta})$ is sufficiently negligible are removed, such that $p(\boldsymbol{x}_0 \mid \boldsymbol{\theta}) \to \mathbb{I}_\Gamma(\boldsymbol{\theta}) \cdot p(\boldsymbol{x}_0 \mid \boldsymbol{\theta})$ remains an accurate approximation for a given target observation $\boldsymbol{x}_0$. Indeed, too small values for the threshold ϵ are inefficient in focusing simulations, whereas too large values would cut into the relevant part of the posterior. For generous enough Γ, i.e. an appropriate threshold ϵ, the posterior remains essentially unchanged far into its tails,

$$\begin{aligned} p_\Gamma(\boldsymbol{\theta} \mid \boldsymbol{x}) &= \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta})\, p_\Gamma(\boldsymbol{\theta})}{\int d\boldsymbol{\theta}\, p(\boldsymbol{x} \mid \boldsymbol{\theta})\, p_\Gamma(\boldsymbol{\theta})} = \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) \cdot \mathbb{I}_\Gamma(\boldsymbol{\theta})\, V^{-1}}{\int d\boldsymbol{\theta}\, p(\boldsymbol{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) \cdot \mathbb{I}_\Gamma(\boldsymbol{\theta})\, V^{-1}} \\ &\simeq \frac{p(\boldsymbol{x} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\boldsymbol{\theta})} = p(\boldsymbol{\theta} \mid \boldsymbol{x})\ . \end{aligned} \tag{2.30}$$

This whole scheme relies on the assumption that posterior estimate $\hat{p}_\Gamma(\boldsymbol{\theta} \mid \boldsymbol{x}_0)$ is a good approximation of $p(\boldsymbol{\theta} \mid \boldsymbol{x}_0)$. An over-confident estimate would remove parameter

ranges that are part of $\Gamma^{(R)}(\boldsymbol{\theta})$. In practice, this effect is not observed for a conservative choice of ϵ. Moreover, one can test the convergence of the sequential scheme by checking whether high probability regions of the estimated posteriors intersect with the boundaries of the indicator function.

2.2.4 Information maximizing neural network

From a practical perspective, a given inference involves training several classifiers (corresponding to the various marginal ratio estimators) in parallel. For complex and high-dimensional data, it is generally useful to compress it before it is input into the inference network through a so-called compression/summary/embedding network. Hence, the inference neural network usually employed to perform TMNRE are generally split into two distinct components: an embedding network $C_\Phi(\boldsymbol{x})$ and binary classification networks $d_\Phi(\boldsymbol{x}, \boldsymbol{\theta})$.

The embedding network learns to compresses the potentially high-dimensional data into a low-dimensional feature vector, estimating the best possible summary statistics from the full input data, $\boldsymbol{s} = C_\Phi(\boldsymbol{x})$. In principle, the compression network can be any sufficiently expressive network, no specialized network architectures are required. In practice, for faster convergence and better results specific inductive biases of different network architectures should be exploited, e.g. for image data, convolutional neural networks are often appropriate. Throughout this thesis, we will see various example of this (e.g. see Chapter 3). Furthermore, the compression network output can be shared as input to all of the classification networks, or subdivided between the classifiers, in order for each compressed summary to be optimized for the specific marginal estimator.

While the purpose of the embedding network is to efficiently summarize data, the purpose of classifiers $d_\Phi(\boldsymbol{s}, \boldsymbol{\theta})$ is to learn the correlation between the data summary $\boldsymbol{s}$ and the parameters of interest $\boldsymbol{\theta}$ in order to perform the actual ratio estimation. Their inputs are the learned summaries of the observational data concatenated with the parameters of interest for the targeted marginal. Usually, the classifiers are parameterized with fully connected layers and unless otherwise specified in this thesis we will be using the ResNet [169] architecture as implemented in `swyft` [57].

The compression network and ratio inference are *optimized simultaneously*, in contrast to other algorithms in the literature, minimizing the BCE loss function (Equation (2.25)). The structure of a general network architecture can be expressed as:

$$d_\Phi(\boldsymbol{s} = C_\Phi(\boldsymbol{x}), \boldsymbol{\theta}) \simeq d_\Phi(\boldsymbol{x}, \boldsymbol{\theta}) = \sigma[\log \hat{r}(\boldsymbol{\theta}; \boldsymbol{x})] \,. \tag{2.31}$$

It is interesting to understand why the approximately equal sign in Equation (2.31) holds and which features $\boldsymbol{s}$ are learned by the embedding network during

training, e.g. in comparison to the hand-crafted ones employed in ABC (Section 2.1). Following Ref. [78], the BCE loss in Equation (2.25) can be written in terms of the Jensen-Shannon divergence (JSD) as

$$\mathcal{L}[d_{\boldsymbol{\Phi}}(\boldsymbol{x},\boldsymbol{\theta})] = 2\log 2 - 2\mathbb{E}_{p(\boldsymbol{x})}\left[D_{JS}(p(\boldsymbol{\theta}\mid\boldsymbol{x})\|p(\boldsymbol{\theta})\right] , \tag{2.32}$$

which follows[6] directly from the definition of the JSD

$$D_{JS}(p\|q) = \frac{1}{2}D_{KL}(p\parallel m) + \frac{1}{2}D_{KL}(q\parallel m) , \tag{2.33}$$

where $m = (p+q)/2$ and D_{KL} denotes the Kullback-Leibler divergence, $D_{KL}(p(x)\mid q(x) = \int \mathrm{d}x p(x)\log\left(\frac{p(x)}{q(x)}\right)$.

We can now apply the same logic to our network with a compression step, an information bottleneck, in Equation (2.31). The posterior is then conditioned on the summary statistics $\boldsymbol{s}$, and the loss function can be written as

$$\mathcal{L}[d_{\boldsymbol{\Phi}}(\boldsymbol{s}=C_\phi(\boldsymbol{x}),\boldsymbol{\theta})] = 2\log 2 - 2\mathbb{E}_{p(\boldsymbol{x})}\left[D_{JS}(p(\boldsymbol{\theta}\mid\boldsymbol{s}=C_\phi(\boldsymbol{x}))\|p(\boldsymbol{\theta})\right] . \tag{2.34}$$

Looking at Equation (2.34), assuming a fully converged classifier and the loss as a function of the summary $\boldsymbol{s}$, it becomes now evident that a data summary $\boldsymbol{s}$ that minimizes the BCE loss has to *maximize* the expected JSD between the data-summary-based posterior and the prior for the parameters of interest $\boldsymbol{\theta}$. Therefore, data summaries $\boldsymbol{s}$ that sufficiently describe data $\boldsymbol{x}$ for a parameter $\boldsymbol{\theta}$ lead to posteriors $p(\boldsymbol{\theta}\mid\boldsymbol{x})$. On the other hand, inefficient data summaries will lead to wider posteriors which are more similar to the prior, reducing the JSD, and increasing the value of the loss function. Hence, it makes sense to expect that the data summaries that are learned are the most informative to discriminate between samples from the posterior and the prior, or, equivalently, to discriminate whether the likelihood-to-evidence ratio is larger or smaller than one,

$$\frac{p(\boldsymbol{s}\mid\boldsymbol{\theta})}{p(\boldsymbol{s})} = \frac{p(\boldsymbol{\theta}\mid\boldsymbol{s})}{p(\boldsymbol{\theta})} \lessgtr 1 . \tag{2.35}$$

[6] In fact,

$$\begin{aligned}
&-2\mathbb{E}_{p(\boldsymbol{x})}\left[D_{JS}\left(p\left(\boldsymbol{\theta}\mid\boldsymbol{x}\right)\parallel p(\boldsymbol{\theta})\right)\right] \\
&\quad= -\mathbb{E}_{p(\boldsymbol{x})}\left[D_{KL}\left(p\left(\boldsymbol{\theta}\mid\boldsymbol{x}\right)\parallel\frac{1}{2}\left(p\left(\boldsymbol{\theta}\mid\boldsymbol{x}\right)+p(\boldsymbol{\theta})\right)\right) + D_{KL}\left(p\left(\boldsymbol{\theta}\right)\parallel\frac{1}{2}\left(p\left(\boldsymbol{\theta}\mid\boldsymbol{x}\right)+p(\boldsymbol{\theta})\right)\right)\right] \\
&\quad= -\int d\boldsymbol{x}\, d\boldsymbol{\theta}\, \underbrace{p(\boldsymbol{x})p(\boldsymbol{\theta}\mid\boldsymbol{x})}_{\to p(\boldsymbol{x},\boldsymbol{\theta})} \ln\frac{p\left(\boldsymbol{\theta}\mid\boldsymbol{x}\right)}{\frac{1}{2}\left(p\left(\boldsymbol{\theta}\mid\boldsymbol{x}\right)+p(\boldsymbol{\theta})\right)} + p(\boldsymbol{x})p(\boldsymbol{\theta})\ln\frac{p\left(\boldsymbol{\theta}\right)}{\frac{1}{2}\left(p\left(\boldsymbol{\theta}\mid\boldsymbol{x}\right)+p(\boldsymbol{\theta})\right)} \\
&\quad= -\int \mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{\theta}\left\{p(\boldsymbol{x},\boldsymbol{\theta})\log d_{\boldsymbol{\Phi}}(\boldsymbol{x},\boldsymbol{\theta}) + p(\boldsymbol{x})p(\boldsymbol{\theta})\log\left[1-d_{\boldsymbol{\Phi}}(\boldsymbol{x},\boldsymbol{\theta})\right]\right\} \\
&\quad= \mathcal{L}[d_{\boldsymbol{\Phi}}(\boldsymbol{x},\boldsymbol{\theta})] - 2\ln 2 ,
\end{aligned}$$

where the third equation follows from Equation (2.19).

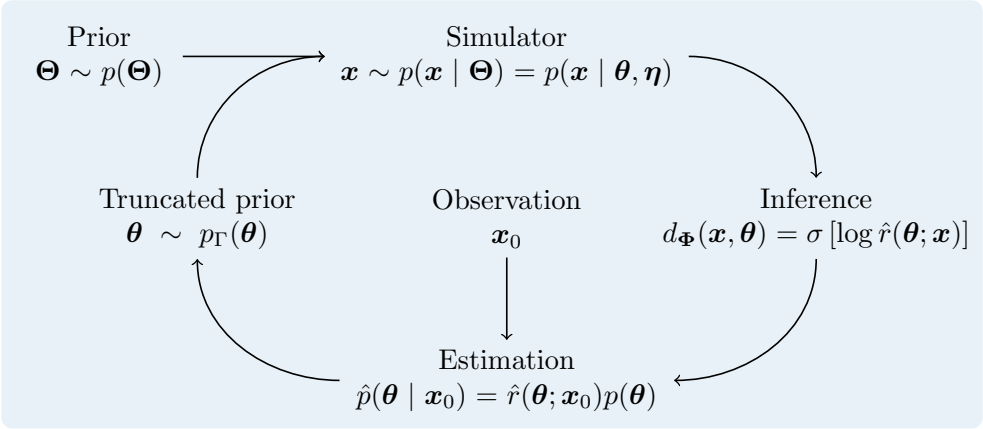

Figure 2.3: **Truncated marginal neural ratio estimation.** Parameters, both of interest $\boldsymbol{\theta}$ and nuisance $\boldsymbol{\eta}$, are sampled from the initial prior $p(\boldsymbol{\Theta})$, and mapped into data $\boldsymbol{x}$ through a programmed simulator that effectively acts as in implicit likelihood. Inference is performed by training binary classifiers $d_\Phi(\boldsymbol{x}, \boldsymbol{\theta})$ with marginal NRE to learn the estimates of the ratios of interest $\hat{r}(\boldsymbol{\theta}; \boldsymbol{x})$ (Sections 2.2.1 and 2.2.2). The ratios are used to weight prior samples, enabling posterior sampling for any observation $\boldsymbol{x}$. Focusing on a target observation $\boldsymbol{x}_0$, we can perform sequential inference (Section 2.2.3) via prior truncation, by constraining the initial parameter space for the inferred parameters to regions $\Gamma^{(R)}(\boldsymbol{\theta})$ that are most relevant for a specific target observation $\boldsymbol{x}_0$. The procedure is repeated until convergence.

In this section we have introduced the technical details of the TMNRE algorithm, and its key building blocks: NRE (Section 2.2.1), marginalization (Section 2.2.2), and prior truncation (Section 2.2.3). We have also seen the general network architecture that is usually employed within this algorithm (Section 2.2.4). A complete sketch of the workflow of a TMNRE algorithm is illustrated in Figure 2.3. We will see in Section 2.3 how to test its results.

2.3 Testing SBI: getting it right

TMNRE is a *locally amortized* technique, meaning that the ratio estimator are trained to learn a function, that can then be evaluated to perform inference very quickly on *any* number of data, thanks to the great evaluation speed of neural networks. This opens up the possibility to perform consistency checks of the statistical properties of the trained networks, within the constrained region over which they have been trained, in each stage of truncation. These type of tests are generally infeasible for

likelihood-based methods, where the algorithms perform inference on a single fixed observation and must rerun from scratch for another observation. In this likelihood-based settings, it is hence computationally very costly to test coverage on simulated data, and usually the statistical properties of the inference results are instead inferred based on convergence criteria of the sampling chains [e.g. 170]. This difference is illustrated in Figure 2.4.

Using local amortization, one can cheaply validate the Bayesian coverage properties of the approximate posteriors [78, 81, 171], and construct confidence regions with exact frequentist coverage [81, 172, 173]. Here we focus on the former types of tests, and use the *expected coverage test* that probes the empirical (i.e. determined from analyses of simulated data $\boldsymbol{x}$) Bayesian coverage properties of the estimated posteriors $\hat{p}(\boldsymbol{\theta} \mid \boldsymbol{x})$.

An *expected coverage test* measures whether Bayesian credible regions of different widths, $\Omega_{p(\boldsymbol{\theta}|\boldsymbol{x})}(1-\alpha)$, have achieved their $1-\alpha$ nominal coverage [78, 171]. In simple words, it checks whether the true parameters $\boldsymbol{\theta}$ fall within the credible region, of the estimated posterior $\hat{p}(\boldsymbol{\theta} \mid \boldsymbol{x})$ for $\alpha\%$ of the randomly drawn observations $\boldsymbol{x} \sim p(\boldsymbol{x} \mid \boldsymbol{\theta})$:

$$1 - \hat{\alpha} \equiv \mathop{\mathbb{E}}_{p(\boldsymbol{x},\boldsymbol{\theta})} \left[\mathbb{I}\left[\boldsymbol{\theta} \in \Omega_{\hat{p}(\boldsymbol{\theta}|\boldsymbol{x})}(1-\alpha) \right] \right] \tag{2.36}$$

$$\underset{\hat{p} \to p}{\to} \mathop{\mathbb{E}}_{p(\boldsymbol{\theta})} \underbrace{\mathop{\mathbb{E}}_{p(\boldsymbol{\theta}|\boldsymbol{x})} \left[\mathbb{I}\left[\boldsymbol{\theta} \in \Omega_{p(\boldsymbol{\theta}|\boldsymbol{x})}(1-\alpha) \right] \right]}_{=1-\alpha} . \tag{2.37}$$

In the limit where the estimator $\hat{p}(\boldsymbol{\theta} \mid \boldsymbol{x})$ approaches the correct one $p(\boldsymbol{\theta} \mid \boldsymbol{x})$, one trivially expect that $1 - \alpha = 1 - \hat{\alpha}$.

In case of a conservative estimator, the nominal credibility $1-\alpha$ is lower than the empirical one $1-\hat{\alpha}$, and vice versa for an overconfident estimator. In general, in case of a conservative coverage plot one should check if the estimator has fully converged. On the other hand, in case of an overconfident coverage plot, one should check for overfitting. In combination with visually checking the posteriors, this test supports the accuracy of the posterior estimator and is also particularly useful when one does not have access to the ground truth against which to compare the results.

Agreement between the nominal and empirically-measured expected coverage is a necessary (but not sufficient) condition for the density/ratio estimator to be a correct estimate of the posterior. Indeed, the expected coverage test checks for the consistency of the posterior, not for its precision. For example, a not focused posterior that resembles the prior leads to a diagonal plot with optimal coverage. Moreover, it tests for the coverage *averaged* over parameter values $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$. As a result, some parameter regions might be over- and other parameter regions under-covered.

Since we are mostly interested in small values of α, corresponding to posterior regions with high mass, it is convenient to reparameterize α in terms of a new variable

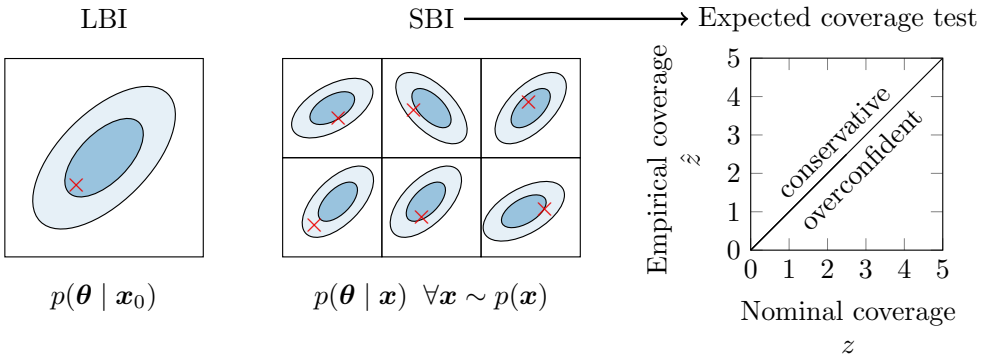

Figure 2.4: Testability: Contrary to LBI algorithms, many SBI methods, including TMNRE, do not estimate just a single posterior, but all of them simultaneously. This property is called "amortization" in the machine learning field [174]. Amortization enables the user to efficiently test the reliability of the inference results, for example with expected Bayesian coverage tests (rightmost panel). The left part of the figure is inspired and adapted from Figure 4 in Ref. [78].

z, defined as $1 - \frac{1}{2}\alpha$ quantile of the standard normal distribution [78]. As a results, $z = 1, 2, 3$ have $1 - \alpha = 0.6827, 0.9545, 0.9997$, and correspond to the common 1σ, 2σ, 3σ regions. We show a mock example of such test in the right panel of in Figure 2.4 to illustrate its general behavior.

In Chapter 3, we will use empirical coverage plots in order to double-check and confirm the convergence of our posterior estimators for strong lensing image analysis. Closely related versions of this test and plots have appeared in the literature, e.g. in Refs. [59, 64, 81]. Developing more tests in order to trust results generated by SBI algorithms is very active and ongoing work [175, 176]. For example, it is worth mentioning an alternative method of defining Bayesian credible regions using distances to random points, which can, in certain circumstances, detect a systematic bias [176].

2.4 Neural SBI methods comparison

Each of the three core neural SBI algorithms described in Section 2.1.2 has its distinct benefits and drawbacks. The most suitable choice between NPE, NLE, and NRE is largely determined by the properties of the underlying problem, the dimensionality of the data and of the parameter space, and the specific application. For a quantitative comparison of the three algorithms on machine learning benchmarks, we refer the reader to Ref. [124]. In this section, following discussions from Refs. [52, 78, 147], we focus on describing qualitative similarities and differences between them.

In general, NPE and NLE solve a density-estimation problem via an *unsupervised* learning task, as there are no explicit labels. Instead, NRE solves a classification problem via a *supervised* learning task, which is often simpler. Moreover, ratios can be estimated with simple classification networks, which are often easier to train than density estimation models (as normalizing flows). Hence, for higher-level inference, learning an estimator for the likelihood-ratio function may often be more efficient.

Furthermore, density estimation becomes increasingly hard when the dimension of the data is large, or the data distribution is highly complex [145]. This problem is alleviated by compressing the data to some simple low-dimensional summary statistics before performing inference with density estimators. For example, a compression network, such as an Information Maximizing Neural Network [177], is often used. The *compression* network is typically trained upfront before inference, with additional simulation cost. On the other hand, as seen in Section 2.2.4, in NRE the compression network and the ratio estimator are trained simultaneously.

One of the key advantages of NPE and NLE over NRE is the additional *sampling functionality* from the surrogate model, which is not possible for NRE parameterized via e.g. a multilayer perceptron (MLP). For NRE and NLE, one needs a way to sample from the neural posterior after training. Still, likelihood(-to-evidence ratio) evaluation is generally very fast thanks to the massive evaluation speed and parallelization afforded by neural networks. Broadly, NPE models are better suited for scenarios requiring repeated inference across multiple observations, as it can be prohibitively expensive to run sampling chains for NRE and NLE for many test points. On the other hand, for hierarchical likelihood sampling, NRE and NLE will be a more effective choice.

Despite these initial differences, the three algorithms share many similarities. First of all, when not using active learning, all three approaches are *amortized*: after an upfront simulation and training phase, they can all be efficiently evaluated for arbitrary data and parameter points (in the prior support). Second, all three approaches can use *active learning*, iteratively updating the proposal distribution to guide the simulator parameters toward the relevant parameter region (Section 2.2.3).

It is also important to note that, while NPE directly targets the posterior, it induces a prior dependence in the inference. On the other hand, NRE and NLE are prior independent, allowing extra flexibility to change the prior during inference, and the possibility to do frequentist inference or model comparisons.

While all approaches can directly target marginal posteriors [120, 160, 165, 178], some extra care is required when combining marginalization and active learning if there is a prior dependence in the inference. Indeed, marginal estimation will be in general affected by the proposal distribution, since one implicitly integrates over it. This generally cannot be undone after training (for a workaround where marginaliza-

tion can be integrated into the compression step by constructing appropriate summary statistics see, e.g., Ref. [178]).

Combining the insights from all algorithms, which are currently categorized as distinct, is expected to lead to facilitate further advancements.

2.5 SBI applications in this thesis

In this first chapter, we have learned about the benefits of simulation-based analysis for data inference, and particularly about the TMNRE algorithm. In the next chapters, we will see its applications to various astrophysical problems: the analysis of strong gravitational lenses as a dark matter probe (Chapter 3), how to scale it to higher dimensional problems with more correlated parameter spaces (Chapter 4), the reconstruction of cosmological initial conditions Chapter 5), and the analysis of a population of point-sources in sky-maps (Chapter 6). The common thread between all these applications is being formally representable by *large and complex forward models*, with many moving parts, on a subpart of which we would like to perform precise scientific inference while correctly propagating uncertainties from the others. This is possible in a SBI framework.

In particular, we will see many of the several types of marginal posteriors that can be directly inferred in simulation-based settings, and which answer specific inference questions relevant for the application at hand. Some of the marginal quantities we will encounter in the different chapters of this thesis are:

- posterior of a single parameter (Chapters 3, 4, and 6);
- 2-dimensional posteriors of parameter pairs (Chapters 3, 4, and 6);
- n-dimensional posteriors of parameters (Chapter 4);
- posterior probability of a pixel value for field reconstruction (Chapter 5);
- posterior probability of an object being present in a specific image region (Chapter 6);
- posterior for the number of objects in an image (Chapter 3 and 6).

Moreover, prior truncation will be achieved by means of different strategies of increasing complexity throughout all of these applications, depending on the their specificities. The main strategies we will see are:

- 1-dimensional box truncation (Chapters 3, 4, and 6);

- n-dimensional hard-likelihood truncation with nested sampling (Chapter 4);
- truncation with tempered likelihood (Chapter 5);
- object detection as prior truncation (Chapter 6).

A detailed contextualization and explanation of the technicalities of each strategy will be given in the respective chapters.

3 Simulation-based inference for strong gravitational lensing

Precision analysis of galaxy-galaxy strong gravitational lensing images provides a unique way of characterizing small-scale dark matter halos, and could allow us to uncover the fundamental properties of dark matter's constituents. Recently, gravitational imaging techniques made it possible to detect a few heavy subhalos. However, gravitational lenses contain numerous subhalos and line-of-sight halos, whose subtle imprint is extremely difficult to detect individually. Existing methods for marginalizing over this large population of sub-threshold perturbers to infer population-level parameters are typically computationally expensive, or require compressing observations into hand-crafted summary statistics, such as a power spectrum of residuals.

In this chapter, we present the first analysis pipeline to combine parametric lensing models and a recently-developed neural simulation-based inference technique called truncated marginal neural ratio estimation (TMNRE) to constrain the warm dark matter halo mass function cutoff scale directly from multiple lensing images. Through a proof-of-concept application to simulated data, we show that our approach enables empirically testable inference of the dark matter cutoff mass through marginalization over a large population of realistic perturbers that would be undetectable on their own, and over lens and source parameters uncertainties. To obtain our results, we combine the signal contained in a set of images with Hubble Space Telescope resolution. Our results suggest that TMNRE can be a powerful approach to put tight constraints on the mass of warm dark matter in the multi-keV regime, which will be relevant both for existing lensing data and in the large sample of lenses that will be delivered by near-future telescopes.

This chapter is based on work from [1].

3.1 Introduction

Dark matter puzzle. Over the past several decades, numerous astrophysical probes including rotational curves of spiral galaxies [179], galaxy-cluster dynamics [180], cosmic microwave background [181], gravitational lensing observations [182], have established dark matter (DM) as one of the major components of the universe, comprising about 85% of its mass. However, up to the present time, the fundamental nature of DM is still one of the key unresolved puzzle in physics. For many years, the cold dark matter (CDM) paradigm has been able to accurately reproduce vastly disparate large-scale observations across all epochs. In this model, DM is massive, neutral, non-relativistic, and collisionless [183].

Despite providing a stunning description of the observed distribution of matter on large scales ($> \mathcal{O}(\text{Mpc})$), the agreement between CDM predictions and observations at galactic and sub-galactic scales has been less clear. At present, there is continued debate over whether the known abundance of dwarf galaxies and the density profiles of low-mass galaxies are in tension with the predictions of ΛCDM (respectively dubbed the missing satellites problem [184, 185] and the cusp-core problem [186], reviewed in Ref. [187]). Solutions to these tensions include the impact of baryonic processes, such as supernovae feedback and reionization processes [188], or alternative DM physics.

The latter approach requires an alteration of DM particle physics, such that large-scale predictions remain unaffected, but the number of small-scale substructures is suppressed. DM models which are warm instead of cold [189–191], collisional instead of collisionless [192], or quantum on macroscopic scales rather than classical [193] predict a diverse array of possible configurations of low-mass halos and could potentially resolve these tensions [194]. Unfortunately, light DM halos are difficult to probe as they are not expected to accumulate enough baryonic matter to form stars and hence are truly *dark* [195, 196]. If DM has significant self-interactions, such halos might be detectable by searching for the self-annihilation or decay products of DM [197]. However, even if such interactions are not present, light halos can potentially be probed through their irreducible gravitational effects. In this chapter we study one such probe: galaxy-galaxy strong gravitational lensing.

Strong gravitational lensing as a dark matter probe. In strong gravitational lensing, the gravitational field of a mass distribution acts as a *lens* by magnifying and distorting the light flux coming from a background *source* [198]. This leads to multiple magnified and distorted images of the source, as explained by general relativity. The effect is sensitive only to how matter is distributed, regardless of its physical nature (baryonic/DM), and thus provides a direct way of probing the distribution of DM at small scales (see Ref. [199] for a recent review).

Indeed, a *perturber* (i.e., a subhalo or line-of-sight (LOS) halo lying somewhere between the observer and source) positioned near one of these images contributes additional, much more localized distortions, on top of the main lens mass distribution. By carefully analyzing the relationship between the multiple images of the source, the distortions from dark perturbers can be disentangled from possible variations in the source light. Thus, a population of dark perturbers can collectively cause perturbations to images that can be detected statistically in order to constrain population-level parameters, such as the suppression scale in the the low-mass end of the halo mass function (HMF), which are dictated by the fundamental properties of DM. Therefore, gravitational lensing provides a pristine probe of small-scale structures and can in principle distinguish between DM scenarios.

Strong lensing image analysis. Various different methods have been suggested to analyse the effects of small-scale structures on lensing images [200]. These methods usually target two different types of lensing systems that differ in the lensed source: quadruply-lensed quasars, and extended background galaxies that get lensed into extended arcs or complete Einstein rings.

In the former case, the source is a nearly point-like quasar that is lensed into four compact images (*"quads"*). These images' positions and flux ratios comprise the summary statistics for these systems. The presence of a perturber near one of these images would cause anomalies in the ratios of their fluxes relative to what would be predicted assuming a smooth lens mass distribution. Evidence for flux ratio anomalies due to perturbers was first found in Ref. [201]. Later, Ref. [31] derived a statistical constraint on the substructure fraction in the lensing galaxies using a small sample of seven lensed quasars. Ref. [202] showed that flux-ratio anomalies can also be used to detect individual low-mass subhalos. Several studies derived upper limits on the subhalo mass function [203], also including perturbations due to LOS halos [33, 204]. Further investigations pointed out the importance of correctly modeling baryonic structure in the main lens, in order to avoid systematic errors while constraining DM substructure abundance with flux-ratio anomalies [34, 205, 206].

Here we focus on *gravitational imaging*, which refers to the analysis of lenses with extended arcs [119, 207, 208]. The observation in this case consists of a whole image. On one hand, such images cover a larger area of the sky than the four point-like images in quads, potentially providing more sensitivity to detect perturbations due to perturbers, in the form of percent-level variations in the shape of the predicted lensed light based on a smooth lens model. On the other hand, extracting this information requires modeling the source galaxy's light, which generally has a complex morphology. The gravitational imaging technique was first introduced in Ref. [209] and further developed in Refs. [119, 208]. Its application to real data has so far yielded to several detections of individual heavy ($> 10^8 M_\odot$) perturbers using deep, high-resolution ob-

servations in the optical from the Hubble Space Telescope (HST) and Keck Telescope as well as in radio data from the Atacama Large Millimeter/sub-millimeter Array [36–38, 210, 211]. Moreover, measurements and non-detections of individual perturbers in samples of gravitational lens systems can be converted to constraints on the (sub)halo mass function and thus dark matter's properties [32, 212, 213].

Established gravitational imaging analyses, such as the method in Refs. [38, 119], use *likelihood-based* inference to infer the properties of perturbers. The central mathematical object in such approaches is the likelihood, a probabilistic model $p(\boldsymbol{x} \mid \boldsymbol{\Theta})$ for the data $\boldsymbol{x}$ given some parameters $\boldsymbol{\Theta}$ for the lens, source, perturbers and possibly other (hyper)parameters.[1] Likelihood-based tools, such as MCMC or nested sampling [214], do not directly produce marginal posteriors but instead compute the *joint posterior* $p(\boldsymbol{\Theta} \mid \boldsymbol{x}_0)$, which must then be marginalized over.

The computational expense of sampling from the high-dimensional joint posterior imposes restrictions on the realism of lensing models that can be analyzed. One such restriction common to most analyses is to assume a particular form of the noise and source model so that the source uncertainties can be excluded from the sampling and marginalized over analytically [36–38, 119, 210]. This makes it difficult to explore more complex source models described by e.g. generative machine learning methods or noise artifacts like cosmic ray streaks that cannot be described by an analytic likelihood.

An additional difficulty with likelihood-based analyses is that each run of MCMC or nested sampling produces posterior samples for just a single observation. Directly exploring the systematics, biases and other statistical properties of a particular lensing model is thus extremely time-consuming, necessitating rerunning posterior sampling many times for different input observations. This also makes analyses such as mapping perturber measurement sensitivity costly. It is noteworthy that recently Ref. [215] pushed to the limits how far one can feasibly go using likelihood-based analyses, fitting 54 images with 5 different mass models.

Likelihood-based analyses also typically assume no more than two perturbers are present in each image. Allowing for n perturbers would cause the joint posterior to become highly multimodal, with approximately $n!$ modes due to exact invariance of the observation under relabeling of perturbers. In these frameworks, it is then challenging to perform statistical inference of quantities such as the posterior for substructure population-level parameter of interest, marginalizing over all source, lens, and substructures parameters to get the marginal of interest.

Transdimensional Bayesian inference partially overcomes this traditional likelihood-based methods' challenge, by using transdimensional MCMC to infer the

[1] For example, the hyperparameters could include the pixel size for pixelated sources or strength of source regularization.

probabilities of different possible populations of perturbers, albeit at substantial computational cost [216, 217].

An alternative strategy involves linearizing the gravitational potential via a *Taylor expansion* of the lens equation. By employing a Taylor expansion, it becomes feasible to capture all small-scales DM substructures without the need to parameterize them directly in the likelihood function [119, 209, 218]. This technique is therefore able to account for the full DM subhalo population (a part from the effects due to the curl-component induced by multi-plane lensing effects).

Another viable approach to reduce the dimensionality of the problem and enable inference of the collective effects of a large number of low-mass substructures at the statistical level requires engineering *summary statistics* from first principles, such as the convergence power spectrum (PS) for different subhalo populations [219–221], and also for LOS populations [222]. However, this approach is not directly applicable to observations, because we do not have access to the true displacement field from the data. Building upon this, it is possible to relate the PS of the surface brightness fluctuations in strong lens images to the lens potential fluctuations arising from DM distribution that contribute to the convergence PS [223–225]. Another summary statistic that has been employed is the residuals between the image and best-fit reconstruction excluding substructures, that is related to the (sub)halo mass function parameters. In particular, this summary statistic has been successfully employed to constrain the HMF suppression scale using ABC [226, 227], a likelihood-free inference method based on a rejection algorithm, as seen in Section 2.1. This approach reduces the dimensionality of the problem and enable inference of the collective effects of a large number of low-mass substructures at the statistical level. However, it is unknown how much information such approach discards.

Another class of methods that has developed in recent years uses *neural networks* to measure lens parameters [228–230], quantifying the structure of gravitational lens potential [231], detect individual subhalos [232], distinguish different types of DM substructure based on their lensing signatures [233], and classify whether each pixel in an image contains a subhalo in a given mass bin [234, 235]. Still, these methods need lots of data to amortize over all possible variations in lensing systems. In fact, amortized methods learn the posterior for any data, generated by any parameter over the whole range of the prior. But learning an amortized posterior is unnecessary if only a small range of parameters are consistent with a target observation.

This work. In this work, we demonstrate that a SBI [52] method called truncated marginal neural ratio estimation [120, 160], from here on TMNRE, can circumvent the inference challenges discussed in the above paragraphs. In particular, we use strong gravitational lensing to constrain warm dark matter (WDM) models [189, 191], which

are well-motivated from a particle physics perspective (see e.g. sterile neutrinos [236] and gravitinos [237]). In WDM models DM particles have non-negligible thermal velocities that allow them to free-stream out of density perturbations, effectively preventing small-scale structure formation. The scale at which this happens depends on model parameters and is parametrised by the half-mode mass, M_{hm}, in the HMF. Therefore, one of the viable ways to discriminate between CDM and alternative WDM models is to constrain the low-mass end of the HMF by probing small-scale DM halos which are completely devoid of stars and truly *dark*, whose only signature is then gravitational.

Here, we present the first analysis pipeline that combines parametric lensing models with recent TMNRE developments to infer the half-mode mass M_{hm} of the subhalo HMF directly from images by combining a set of realistic simulated galaxy-galaxy strong lenses. In fact, there are currently around a hundred strong lensing observations suitable for substructure inference, most of which come from the SLACS [238] and BELLS [239] surveys. In the near future, new and future telescopes like JWST [12], ELT [11], Euclid [13, 42], SKA [240], and LSST [241] will greatly increase the quality of data suitable for gravitational imaging analyses as well as its quantity, from $\mathcal{O}(100)$ to $\mathcal{O}(10^5)$ images [242, 243]. It is then extremely important to be able to combine the information coming from different observations in the statistical analysis.

As discussed in Chapter 2, SBI refers to a class of statistical inference methods that use the output of a stochastic simulator that need not have a known likelihood. In particular, NRE, first presented in Ref. [153], trains a neural network to map from observations directly to *marginal posteriors* for a specified subset of model parameters (e.g. the Einstein radius of a lens). This bypasses the requirement of likelihood-based inference to sample the joint posterior. In contrast to methods like ABC, this also removes the need to engineer summary statistics [227] as they are in effect learned directly from the training data. Since NRE learns a marginal posterior *function*, it is straightforward to check the statistical properties of the inference results for different observations. TMNRE further extends NRE by focusing training data generation in the regions of parameter space most relevant for analyzing a particular observation over a sequence of inference rounds. This substantially reduces the number of simulations required to train the inference network as well as the required network complexity. This truncation method applied to strong-lensing images was proposed in Ref. [41] and used in Ref. [244] to learn marginal posterior approximations for individual subhalo parameters, marginalizing over lens and source uncertainties given an observation.

Several other works have applied SBI to substructure lensing. In Ref. [71] a likelihood-ratio estimation technique similar to TMNRE was employed to measure density profile parameters of subhalos from images. Ref. [68] recently applied neural posterior estimation to measure the subhalo HMF normalization in mock lensing

images using real galaxy images as sources. Ref. [70] utilized a "likelihood-based" SBI method requiring the simulator's score[2] to measure the slope and normalization of a subhalo HMF in simple mock images.

The present work complements these efforts in several ways. First, it demonstrates that our *hierarchical* TMNRE approach is able to efficiently and accurately infer the statistic of the HMF suppression scale given a set of HST resolution observation. Then, it illustrates the importance of combining the information coming from different observations in the statistical analysis, especially in light of near-future data delivery. Lastly, it shows that TMNRE enables statistical checks that would be extremely expensive with likelihood-based inference since they would require rerunning the sampling machinery on numerous mock observations.

This chapter is organized as follows. In Section 3.2 we describe our strong lensing model, which uses an analytic source and main lens in conjunction with a well-motivated perturber model that accounts for both subhalos and LOS halos. In Section 3.3 we briefly discuss the inference methodology based on TMNRE employed in the statistical analysis. Then, we show our results for hierarchical inference of the halo mass function suppression scale from an ensemble of lenses in Section 3.4. Finally, we conclude in Section 3.6. This work will help form the basis for SBI-based analysis of strong lensing images as a dark matter probe in existing and future lensing data.

3.2 Modeling strong lensing observations

Here we review how we model strong lensing images. We implement our lensing model in `PyTorch` [109] so that we can leverage GPUs to rapidly generate large numbers of observational data (a single-channel telescope image).

3.2.1 The physics of strong lensing

Before delving into modeling details, we briefly summarize the key points of the physics of gravitational lensing, referring the reader to e.g. Ref. [27] for a more detailed overview. In strong-lensing systems the mass distribution of a foreground galaxy gravitationally lenses the light rays coming from a background source, resulting in an arc-like image in the case of an extended galaxy source. We assume that mass densities are low enough to treat the gravitational field of the matter in the image plane in the Newtonian approximation of general relativity (GR). In this case the metric is fully characterized by the lens' gravitational potential ψ. We also adopt the

[2]The score is the derivative of the log-likelihood for a given observation with respect to the model's parameters.

thin lens approximation, which assumes all the lens mass lies in a single *image plane* and all the source light is emitted from a *source plane*.

We will be using $\vec{\xi} \equiv (\xi_x, \xi_y)$ and $\vec{x} \equiv (x, y)$ as two-dimensional angular coordinates in the image and source planes respectively, and use z to indicate distances along the orthogonal dimension. Since the image plane covers a small angular patch of the sky and the lensing deflections are small in the Newtonian limit, the coordinate system can be treated as Cartesian.

The configuration, then, is determined by two fields: the distribution of surface brightness in the source plane, $\beta(\vec{x})$, and the distribution of mass in the image plane, described by the projected potential:

$$\Psi(\vec{\xi}) \equiv \int_{-\infty}^{\infty} \psi(\xi_x, \xi_y, z) \mathrm{d}z \tag{3.1}$$

Under these assumptions, the source-plane coordinate to which a light ray through the image plane traces back is given by the simple *lens equation*

$$\vec{x} = \vec{\xi} - \vec{\alpha}(\vec{\xi})\,. \tag{3.2}$$

The displacement field $\vec{\alpha}$ is determined by the projected potential:

$$\vec{\alpha}(\vec{\xi}) = \frac{D_{LS}}{D_S} \frac{2}{c^2} \frac{\vec{\nabla}_{\vec{\xi}} \Psi(\vec{\xi})}{D_L}\,, \tag{3.3}$$

where the gradient is taken in the image plane and should have dimensions of inverse length, whence the introduction of the observer-lens angular diameter distance D_L. The expression also involves the angular diameter distances D_{LS} (from the lens to the source), and D_S (from the observer to the source).[3] We illustrate in Figure 3.1 the geometry of the system.

From a computational standpoint it is more convenient to represent Equation (3.3) as an integral. Using the Poisson equation $\nabla^2 \Psi = 4\pi G \rho$ and setting appropriate boundary conditions:

$$\vec{\alpha}(\vec{\xi}) = \frac{4G}{c^2} \frac{D_{LS}}{D_L\, D_S} \int \Sigma(\vec{\xi}') \frac{\vec{\xi} - \vec{\xi}'}{|\,\vec{\xi} - \vec{\xi}'\,|^2} \mathrm{d}^2(D_L \vec{\xi}')\,, \tag{3.4}$$

where the integral is performed over the image plane, and where c is the speed of light and G is the gravitational constant.

Here Σ is the projected (surface) mass density related to the 3D mass density ρ by an integration along the coordinate perpendicular to the lens plane z:

$$\Sigma(\vec{\xi}) \equiv \int_{-\infty}^{\infty} \rho(\xi_x, \xi_y, z) \mathrm{d}z, \tag{3.5}$$

[3] We compute angular diameter distances with `astropy` [245, 246] using the flat cosmology from Planck [247].

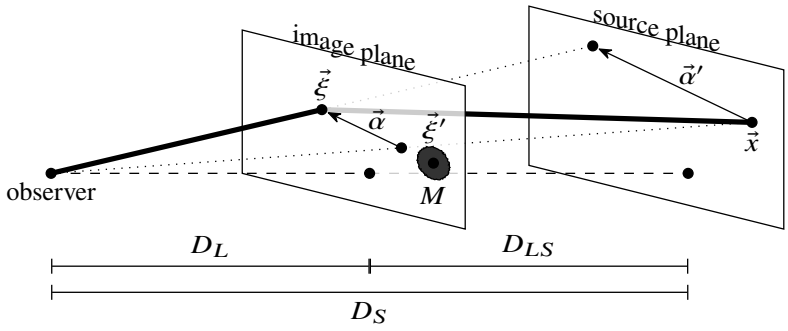

Figure 3.1: Thin lens geometry. The lensing mass M located at $\vec{\xi}'$ bends the light ray (thick line) emanating from the point $\vec{x}$, so that to the observer it looks like it is coming from the direction of $\vec{\xi}$. General relativity predicts the deflection angle $\vec{\alpha}'$ *as viewed from the image plane* based on the mass M and the impact parameter $\vec{b} = \vec{\xi} - \vec{\xi}'$. It then has to be rescaled by $\frac{D_{LS}}{D_S}$ to obtain the displacement field $\vec{\alpha}$ (as viewed by the observer) for use in Equation (3.2). Angles are all assumed to be small enough that they can be used for Euclidean calculations. The dashed line is the optical axis perpendicular to the planes and connects the origins of the coordinate systems for each plane. The figure is a reproduction of Figure 2 in Ref. [41].

similarly to the expression for the projected potential.

The expression in Equation (3.4) can be simplified by introducing the convergence κ in terms of the critical surface density Σ_{cr}:

$$\kappa(\vec{\xi}) = \frac{\Sigma(\vec{\xi})}{\Sigma_{\text{cr}}}, \quad \Sigma_{\text{cr}} \equiv \frac{c^2}{4\pi G}\frac{D_S}{D_L D_{LS}}. \tag{3.6}$$

If $\kappa > 1$, a lens can form multiple images [27, Section 2.6]. The threshold for this to happen is set by the critical surface density, and it is evident from its expression that for a fixed distance to the lens (D_L) further away sources (D_S) are lensed more easily, since they require smaller deflections.

Furthermore, it can be shown that the convergence κ is related to the isotropic part of the trace of the Jacobian of the lensing transformation through

$$\frac{\mathrm{d}\vec{x}}{\mathrm{d}\vec{\xi}} = \begin{pmatrix} 1-\kappa & 0 \\ 0 & 1-\kappa \end{pmatrix} - \begin{pmatrix} \gamma_1 & \gamma_2 \\ \gamma_2 & -\gamma_1 \end{pmatrix}, \tag{3.7}$$

where the second part is the shear (discussed more in details in Section 3.2.3). Another important quantity is the inverse of the Jacobian's determinant, the magnification

$$\mathcal{M} \equiv \left| \frac{\mathrm{d}\vec{\xi}}{\mathrm{d}\vec{x}} \right| = \left[(1-\kappa)^2 - \left(\gamma_1^2 + \gamma_2^2\right) \right]^{-1}, \tag{3.8}$$

which is a measure of how much the solid angle spanned by the source is enlarged, or equivalently, of how gravitational focusing directs a larger fraction of the energy radiated by the source to the observer [248]. Thus, when the convergence κ and the shear γ are equal or greater than unity, sources are *strongly* magnified, and we can use the magnitude of the magnification to differentiate between *strong* from *weak* lensing regimes.

While lensing does change the apparent solid angle of a source, it is worth noting that it conserves energy, since it merely alters the trajectories of photons rather than creating or destroying them. As a result, the surface brightness $\beta(\vec{\xi})$ in the image plane is equal to the surface brightness at the point to which it traces back in the source plane (assuming one isolated source):

$$\beta(\vec{\xi}) = \beta(\vec{x}(\vec{\xi}))\,. \tag{3.9}$$

Given the physics of strong lensing, in order to fully specify a strong-lensing model we then need two main ingredients: the lens model, which describes the total mass distribution of the lens, and the source model, which describes the surface brightness profile of the background source. It is common to split the lens model into a macroscopic smooth component (main lens and external shear) and a substructure[4] component, due to subhalos and line-of-sight halos. Each lensing ingredient can be directly superimposed by summing their respective displacement fields in the lens plane:

$$\vec{\alpha} = \vec{\alpha}_{\rm lens} + \vec{\alpha}_{\rm ext} + \sum_{i=1}^{N_{\rm sub}} \vec{\alpha}_{\rm sub,i} + \sum_{i=1}^{N_{\rm los}} \vec{\alpha}_{\rm los,i}. \tag{3.10}$$

In the following sections, we will describe each component of the model we use to simulate mock images of gravitational lenses: the source, the main lens, the dark matter perturbers, and instrumental noise. The model and its priors are summarized in Table 3.1.

3.2.2 Source model

To model the surface brightness of the source galaxy, we adopt the widely-used Sérsic profile [251]. The surface brightness distribution is parameterized by

$$\beta(\vec{x}) = I_e \exp\left\{-k_n \left[\left(\frac{R(\vec{x})}{r_e}\right)^{1/n} - 1\right]\right\}, \tag{3.11}$$

[4]Throughout this thesis, we use the terms 'small-scale structures', 'substructures', and 'low-mass halos' when considering both subhalos of the main lens and line-of-sight halos.

Parameter	True value	Prior	Description
Main lens			SPLE
θ_E [$''$]		$\mathcal{U}(1., 2.)$	Einstein radius
$\xi_{0,x}$ [$''$]		$\mathcal{U}(-0.2, 0.2)$	lens center x axis
$\xi_{0,y}$ [$''$]		$\mathcal{U}(-0.2, 0.2)$	lens center y axis
q_l		$\mathcal{U}(0.1, 1.)$	axis ratio
ϕ_l [rad]		$\mathcal{U}(0, 2\pi)$	rotation angle
γ	2.1	-	slope
$z_{\rm lens}$	0.5	-	lens redshift
External shear			
γ_1		$\mathcal{U}(-0.05, 0.05)$	$1^{\rm st}$ component
γ_2		$\mathcal{U}(-0.05, 0.05)$	$2^{\rm nd}$ component
Source			Sérsic
I_e		$\mathcal{U}(0., 4.)$	surface intensity
r_e [$''$]		$\mathcal{U}(0.1, 2.5)$	effective radius
x_0 [$''$]		$\mathcal{U}(-0.1, 0.1)$	source center x axis
y_0 [$''$]		$\mathcal{U}(-0.1, 0.1)$	source center y axis
q_s		$\mathcal{U}(0.1, 1.)$	axis ratio
ϕ_s [rad]		$\mathcal{U}(0, 2\pi)$	position angle
n		$\mathcal{U}(0.1, 4.)$	index
$z_{\rm src}$	2	-	source redshift
Subhalos			tNFW
$\vec{p}$ [$''$]	$\in [-2.5, 2.5]$	$\mathcal{U}_{\rm 2D}(-2.5, 2.5)$	position
m_{200} [$\rm M_\odot$]	$\in [10^7, 10^{10}]$	[249]	virial mass
c_{200}	15.	-	concentration
τ	6.	-	truncation
LOS halos			projected tNFW
$\vec{p}$ [$''$]	$\in [-2.5, 2.5]$	$\mathcal{U}_{\rm 2D}(-2.5, 2.5)$	position
m_{200} [$\rm M_\odot$]	$\in [10^7, 10^{10}]$	[250]	virial mass
$z_{\rm LOS}$	$\in [0, z_{\rm src}]$	[250]	LOS redshift
c_{200}	15.	-	concentration
τ	6.	-	truncation
WDM			
$\rm M_{hm}$ [$\rm M_\odot$]		$\log \mathcal{U}(10^7, 10^{10})$	half-mode mass

Table 3.1: Summary of model parameters used for the simulated images in this work. When a prior distribution is not specified, the parameter is fixed to the true value.

where I_e is the surface intensity at the half-light radius r_e. The radial parameter $r(\vec{x}) = \sqrt{r_x^2 + r_y^2}$ is the length of the elliptical radial coordinate vector

$$\begin{pmatrix} r_x \\ r_y \end{pmatrix} = \begin{pmatrix} \sqrt{q_s} & 0 \\ 0 & 1/\sqrt{q_s} \end{pmatrix} \begin{pmatrix} \cos\phi_s & \sin\phi_s \\ -\sin\phi_s & \cos\phi_s \end{pmatrix} \begin{pmatrix} x - x_{0,s} \\ y - y_{0,s} \end{pmatrix}, \tag{3.12}$$

which depends on the source's position angle ϕ_s, axis ratio q_s, and center of light position $(x_{0,s}, y_{0,s})$.

The normalization k_n is related to the index n by an implicit transcendental equation in terms of the complete and lower incomplete gamma functions $2\gamma(2n, k_n) = \Gamma(2n)$. We use the expansion in series from Ref. [252],

$$k_n \approx 2n - \frac{1}{3} + \frac{4}{405n} + \frac{46}{25515n^2} + \frac{131}{1148175n^3} - \frac{2194697}{30690717750n^4}\,. \tag{3.13}$$

valid over a wide range of indices, $n > 0.36$. For typical galaxies $1/2 < n < 10$.

Therefore, our source model is parametrised by seven variables that we collect in the vector $\boldsymbol{\eta}_{\rm src} \equiv \{I_e, r_e, x_{0,s}, y_{0,s}, q_s, \phi_s, n\}$. We fix the source's redshift to $z_{\rm source} = 2$.

3.2.3 Main lens model

We adopt the singular power law ellipsoid (SPLE) model for the main lens galaxy, which is capable of modeling the gravitational potential of strong lenses to near the percent level [253]. Furthermore, the SPLE model has been shown to be a more adequate representation of the combined DM and baryon mass distribution in the inner regions of galaxies, to which strong lensing is most sensitive [253]. The SPLE deflection field can be expressed in closed-form as a complex field $\alpha = \alpha_x + i\alpha_y$ [254, 255]:

$$\begin{aligned} \vec{\alpha}^{\rm SPLE}(\vec{\xi}) = \theta_E \frac{2q_{\rm l}^{1/2}}{1+q_{\rm l}} \left(\frac{\theta_E}{r}\right)^{\gamma-2} e^{i\varphi} \\ \cdot {}_2{\rm F}_1\left(1, \frac{\gamma-1}{2}, \frac{5-\gamma}{2}, -\frac{1-q_{\rm l}}{1+q_{\rm l}} e^{2i\varphi}\right). \end{aligned} \tag{3.14}$$

Here (r, φ) are elliptical coordinates, related to the Cartesian coordinates $\vec{\xi}$ through a transformation parametrized by the lens' orientation $\phi_{\rm l}$, axis ratio $q_{\rm l}$ and position $(\xi_{\rm x,0}, \xi_{\rm y,0})$:

$$\begin{pmatrix} r_x \\ r_y \end{pmatrix} = \begin{pmatrix} q_{\rm l}^{1/2} & 0 \\ 0 & q_{\rm l}^{-1/2} \end{pmatrix} \begin{pmatrix} \cos\phi_{\rm l} & \sin\phi_{\rm l} \\ -\sin\phi_{\rm l} & \cos\phi_{\rm l} \end{pmatrix} \begin{pmatrix} \xi_x - \xi_{x,0} \\ \xi_y - \xi_{y,0} \end{pmatrix}, \tag{3.15}$$

$$\tan\varphi = \frac{r_y}{r_x}. \tag{3.16}$$

In the circular ($q = 1$) isothermal ($\gamma = 2$) case this reduces to $\alpha = \theta_E e^{i\varphi}$, i.e. is an isotropic constant, as is well-known. Since the hypergeometric function ${}_2\mathrm{F}_1$, however, is not implemented in `PyTorch`, we instead pretabulate its value as a function of ϕ_l, q_l and γ and interpolate at runtime, as described in Ref. [256]. This results in very fast code with negligible output degradation.

The slope γ has a complicated degeneracy with the size of the source [257, 258]. Roughly, larger γ values cause the spatial scale of the source to increase [35, Section 3.3]. For simplicity we fix $\gamma = 2.1$. In principle, inferring the slope is possible, but it requires more training data and leads to increased uncertainties in both lens and source parameters.

We also assume the lens galaxy's light has been perfectly subtracted, and fix its redshift to $z_\mathrm{lens} = 0.5$.

To account for the weak lensing due to large-scale structure located along the line of sight to the source, we also include an external shear component, which is constant across the image plane:

$$\vec{\alpha}^\mathrm{shear}(\vec{\xi}) = \begin{pmatrix} \gamma_1 & \gamma_2 \\ \gamma_2 & -\gamma_1 \end{pmatrix} \vec{\xi}. \tag{3.17}$$

Our main lens model thus has seven parameters: the SPLE parameters $(\xi_{x,0}, \xi_{y,0}, \phi_\mathrm{l}, q_\mathrm{l}, \theta_E)$ and the external shear parameters (γ_1, γ_2), which we denote collectively with $\boldsymbol{\eta}_\mathrm{lens}$.

3.2.4 Small-scale structures model

DM substructures can be divided into two categories: subhalos that orbit around the main halo at the lens redshift, and LOS halos distributed between the source and the observer. LOS halos are a more direct probe of free-streaming-induced small-scale structure suppression, because they are less affected by baryonic processes and environmental effects, such tidal stripping interactions with the main halo [259]. For this reason and the fact that they are expected to be more abundant than subhalos in a lensing system [259, 260], it is very important to model them as well, in order to correctly estimate the collective effects of all substructures on the lensing image.

Density profile

To model the density profiles of small-scale DM halos we adopt the smoothly truncated universal 3D mass density profile from Ref. [261]:

$$\rho_\mathrm{tNFW}(r) = \frac{\rho_s}{r/r_s(1 + r/r_s)^2} \frac{1}{1 + (r/r_t)^2}. \tag{3.18}$$

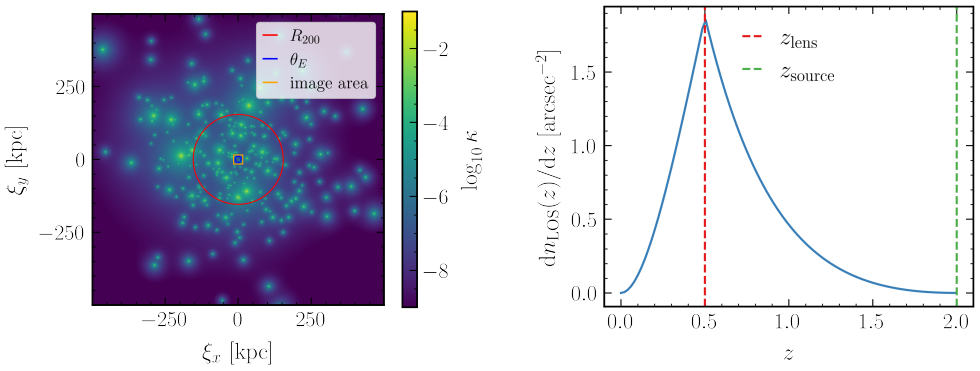

Figure 3.2: *Left panel:* Convergence map for a CDM subhalo population in the adopted mass range. The convergence map shows how the deflecting mass from all the subhalo lenses is distributed. The full map size is 1×1 Mpc. We mark in red the virial radius of the main lens halo, in blue its Einstein radius, and in orange the 5×5 arcsec lensing image area. *Right panel:* LOS halos distribution in redshift for our source and lens redshifts configuration, $z_{\rm lens} = 0.5$ and $z_{\rm source} = 2$.

Here r is the three-dimensional distance from the center of the halo, ρ_s and r_s are respectively the scale density and scale radius that specify a Navarro-Frenk-White (NFW) profile [262], and $r_t \equiv \tau r_s$ is the tidal truncation radius that depends on the history of the subhalo. Typical values of the truncation scale τ range from $4-10$ for spherically symmetric lenses [33, 224]; we fix $\tau = 6$ for simplicity. Compared to the standard NFW form, which has an infinite total mass, the truncated NFW contains an additional truncation term that makes the profile decay as r^{-5} for large radii, resulting in a finite total mass given by:

$$m_\tau = 4\pi\rho_s r_s^3 \frac{\tau^2}{(\tau^2+1)^2}[(\tau^2-1)\ln\tau + \tau\pi - (\tau^2+1)]. \tag{3.19}$$

With a fixed truncation scale, the truncated NFW profile is fully determined by the same parameters that determine the NFW profile: the virial mass $m_{200,\rm sub}$[5] and the concentration $c = r_{200}/r_s$ of the halo. The latter measures how concentrated the mass of a halo is and fixes the density normalization; in principle, it varies from one subhalo to the next and shows dependencies on mass and redshift of the main halo. Here, instead of adopting a concentration-mass relation, we fix $c_{200} = 15$, which is roughly the average value for perturbers in the mass range $10^7 - 10^{10}\mathrm{M}_\odot$ [263, Figure

[5]We parameterize subhalos by what would be their mass up to the virial radius r_{200} using the untruncated profile, with the same central density ρ_s and scale radius r_s as the truncated one. This is defined as the mass of the halo enclosed in a sphere where the untruncated halo's average density is 200 times the critical density.

7]. We anticipate that accounting for scatter in the mass-concentration relation might actually improve our ability to measure subhalos' parameters as higher concentrations lead to substantially stronger lensing signals [264]. The equations for calculating the displacement field of a truncated NFW halo, given its mass and position, are fully elaborated in Appendix A of Ref. [261].

To generate perturber populations, we must choose values for their density normalizations and scale radii. Since simulation studies typically measure the halo mass $m_{200,\mathrm{sub}}$ and the concentration c, it is more convenient to sample populations from distributions over these parameters. These variables can then be mapped to the parametrization above via

$$r_s = \frac{1}{c}\left[\frac{3m_{200,\mathrm{sub}}}{4\pi\, 200\rho_{\mathrm{cr}}(z_{\mathrm{lens}})}\right]^{1/3}, \tag{3.20}$$

$$\rho_s = \rho_{\mathrm{cr}}(z_{\mathrm{lens}})\,\frac{1}{3}\frac{c^3}{\log(1+c) - c/(1+c)}\,. \tag{3.21}$$

For simplicity, we model LOS halos using exactly the same profile even though they typically have not undergone tidal stripping.

The parameters of an individual subhalo which are not fixed are thus $\boldsymbol{\eta}_{\mathrm{sub}} \equiv (x_{\mathrm{sub}}, y_{\mathrm{sub}}, m_{200,\mathrm{sub}})$, where the first and second components are the projected position of the subhalo. In the case of LOS halos, the parameter set also includes the redshift z_{los}. In the following sections, we will use arrows to denote vectors, e.g. $\vec{m}_{200,\mathrm{sub}}$ is an ordered set of masses, one for each simulated subhalo, and bold letters with an arrow to indicate arrays of vectors, e.g. $\vec{\boldsymbol{\eta}}_{\mathrm{sub}} \equiv (\boldsymbol{x}_{\mathrm{sub}}, \boldsymbol{y}_{\mathrm{sub}}, \boldsymbol{m}_{200,\mathrm{sub}})$ is an ordered set of positions and mass in the lens plane, one for each simulated subhalo, and equivalently for LOS halos $\vec{\boldsymbol{\eta}}_{\mathrm{los}} \equiv (\boldsymbol{x}_{\mathrm{los}}, \boldsymbol{y}_{\mathrm{los}}, \boldsymbol{z}_{\mathrm{los}}, \boldsymbol{m}_{200,\mathrm{los}})$. We collectively define the parameters of the full population of perturbers, subhalos and LOS halos, as $\vec{\boldsymbol{\eta}}_{\mathbf{p}}$. In the next two subsections, we describe how we sample these parameters.

Generating subhalos

We sample subhalo masses from the CDM mass function of Ref. [249]:

$$\frac{1}{m_{200}}\frac{\mathrm{d}n_{\mathrm{sub}}(m_{200}, z)}{\mathrm{d}\log m_{200}} = A_M(1+z)^{1/2} m_{200}^{\alpha} \exp\left[-\beta\left(\frac{m_{200}}{M}\right)^3\right], \tag{3.22}$$

where M [6] is the mass of the main lensing halo and m_{200} the subhalo mass. The free parameters in this function were fit to hydrodynamical cosmological simulations that included baryons in Ref. [265]. In particular, we use the fits to EAGLE, which give $\alpha = 0.85$ (given in the text) and $(A_M, \beta) = (2.4 \times 10^{-4}\,\mathrm{M}_\odot{}^{\alpha-1}, 300)$ (extracted from their figures). The expected number of subhalos in a given mass interval for the lens halo system can be computed by integrating the mass function over that interval.

The spatial distribution of subhalos has been shown to follow an Einasto profile [266]. However, since the virial radius of a typical main lens halo is much larger than its Einstein radius, and hence, than the image plane, we approximate the distribution to be uniform in the lensing image area. Still, we derive the total number of expected subhalos within the image, $\bar{n}_{\mathrm{sub}}$, via the Einasto fit of Ref. [265]. We find that on average $\bar{n}_{\mathrm{sub}} = 4$ subhalos fall within the lensing image area in our adopted lensing configuration and mass range. Thereafter we generate the subhalo population by sampling the number of subhalos from Poisson($\bar{n}_{\mathrm{sub}}$), drawing their masses from the subhalo mass function in Equation (3.22) and sampling their projected positions uniformly over the lens plane. Since the vast majority of subhalos fall outside the lens plane, we expect their lensing effect to be mostly degenerate with external shear, and thus do not simulate them. In the left panel of Figure 3.2 we show the convergence map for one realization of our subhalo population.

Generating line-of-sight halos

As described in Ref. [222], we first compute the average number of LOS halos in the double-pyramid geometry connecting the observer, lens-plane and source. For each simulation we sample the number of LOS halos from Poisson($\bar{n}_{\mathrm{los}}$). We then sample their redshifts and projected positions uniformly over the double-pyramid region and draw their masses from the mass function in Ref. [250], assuming an overdensity with respect to the critical density of the universe at the epoch of analysis of $\Delta = 200$.

More in details, we infer the number of detectable LOS halos by integrating their mass function in the mass range adopted for the analysis and within the double-cone volume

$$\bar{n}_{\mathrm{los}} = \int_0^{z_{\mathrm{src}}} \int_{m_{200,\mathrm{min}}}^{m_{200,\mathrm{max}}} n_{\mathrm{los}}(m_{200}, z)\, \mathrm{d}m_{200}\, \frac{\mathrm{d}V}{\mathrm{d}z} \mathrm{d}z. \tag{3.23}$$

On average we get $\bar{n}_{\mathrm{los}} = 260$ LOS halos projected in our lens plane. Similarly to what we do with the subhalo population, when generating simulated images, we draw the

[6] The total mass of the lens galaxy is described by the Einstein radius of the system, a very well-constrained parameter in lensing inference analyses. For the purpose of describing subhalos, we need to be able to map the measured properties of the lens (the Einstein radius θ_{E}) onto the properties of the host halo (the mass M). For simplicity, we compute the mass of the host halo transforming the Einstein radius distance measure into a mass measure. We would like to point out a similar approach from Ref. [70], where they relate the central velocity dispersion of a singular isothermal ellipsoid lens mass distribution profile to the virial mass of the host halo.

number of LOS halos from Poisson($\bar{n}_{\rm los}$), we then sample their masses and redshift from the halo mass function and sample their projected positions uniformly over the lensing image area. As with subhalos, we ignore any LOS halos lying outside the double pyramid volume. In right panel of Figure 3.2 we show the distribution of LOS halos in redshift for our lens and source redshifts configuration.

To avoid expensive iterative ray-tracing through the lens planes of each LOS halo, we project them as effective subhalos into the lens plane, using the relations derived in Ref. [222] to rescale their scale radii and masses. Following Ref. [222], LOS halos at comoving distance χ can be treated as subhalos on the main-lens plane with an effective projected mass density given by:

$$\Sigma_{\chi,\rm eff}(D_L\vec{x}; m_{200}, r_s, \tau) = \Sigma(D_L\vec{x}; m_{200,\rm eff}, r_{s,\rm eff}, \tau). \tag{3.24}$$

The effective scale radius $r_{s,\rm eff}$ and mass $m_{200,\rm eff}$ are respectively

$$r_{s,\rm eff} = \frac{D_L}{g(\chi)D_\chi} r_s, \tag{3.25}$$

and

$$m_{200,\rm eff} = f(\chi)\frac{\Sigma_{\rm cr,L}}{\Sigma_{\rm cr,\chi}}\left(\frac{D_L}{g(\chi)D_\chi}\right)^2 m_{200}. \tag{3.26}$$

The piecewise functions $f(\chi)$ and $g(\chi)$ are:

$$f(\chi) = \begin{cases} 1-\beta_{\chi L} & \chi \le \chi_L \\ 1-\beta_{L\chi} & \chi > \chi_L \end{cases}, \tag{3.27}$$

and

$$g(\chi) = \begin{cases} 1 & \chi \le \chi_L \\ 1-\beta_{L\chi} & \chi > \chi_L \end{cases}, \tag{3.28}$$

with $\beta_{ij} = \dfrac{D_{ij}D_s}{D_j D_{is}}$, where D_i is the angular diameter distance from the observer to plane i, and D_{ij} is the angular diameter distance from lens plane i to lens plane j, and χ_L is the comoving distance to the main-lens plane. We have also introduced the critical surface density at plane i

$$\Sigma_{{\rm cr},i} \equiv \frac{c^2 D_s}{4\pi G D_i D_{is}}. \tag{3.29}$$

It should be noted that effective convergence methods, like the one we adopt [222], do not fully capture the subtleties and degeneracies of multi-plane lens analysis, by disregarding how, when a DM small-scale halo is not in the lens plane, the lens mass

model can absorb its lensing signal [260, 264, 267]. The omission of this effect may lead to an overestimate of the LOS halos contribution and needs to be addressed before this technique can be safely used for the analysis of real data.

For LOS halos we adopt the same concentration and truncation scale values used for subhalos.

3.2.5 Modeling free-streaming effects in WDM

In WDM models DM particles have non-negligible thermal velocities that allow them to free-stream out of density perturbations, effectively preventing small-scale structure formation. The free-streaming effects of WDM are well described in terms of the half-mode wavelength $\lambda_{\rm hm}$, which corresponds to the scale at which the DM transfer function falls to half the CDM transfer function. We can define the half-mode mass as the mass contained within a radius of the half-mode wavelength:

$$\mathrm{M_{hm}} = \frac{4\pi\Omega_{\rm M}\rho_{\rm crit}}{3}\left(\frac{\lambda_{\rm hm}}{2}\right)^3, \tag{3.30}$$

where $\Omega_{\rm M}$ is the matter density parameter and $\rho_{\rm crit}$ is the critical density of the universe. Following Ref. [268], the half-mode wavelength,

$$\lambda_{\rm hm} = 2\pi\alpha_{\rm hm}\left(2^{\nu/5}-1\right)^{-1/(2\nu)}, \tag{3.31}$$

is the scale below which the initial density perturbations are completely erased, with $\nu = 1.12$ and, assuming that all DM is warm,

$$\alpha_{\rm hm} = 0.049\left(\frac{m_{\rm WDM}}{\rm keV}\right)^{-1.11}\left(\frac{\Omega_{\rm DM}}{0.025}\right)^{0.11}\left(\frac{h}{0.7}\right)^{1.22} h^{-1}\mathrm{Mpc}. \tag{3.32}$$

We then have a one-to-one mapping between the mass of the WDM particle and the half-mode mass. For strong lensing, the half-mode mass can be thought of as an effective cutoff mass below which the DM mass function is strongly suppressed. To model this suppression in the WDM mass function we adopt for both subhalos and LOS halos the functional form from Ref. [269]:

$$\frac{n_{\rm WDM}}{n_{\rm CDM}} = \left(1+\left(\alpha\frac{\mathrm{M_{hm}}}{m_{200}}\right)^{\beta}\right)^{\gamma}, \tag{3.33}$$

with best fit parameters $\alpha = 4.2$, $\beta = 2.5$, and $\gamma = 0.2$ for subhalos, $\alpha = 2.3$, $\beta = 0.8$, and $\gamma = 1$ for central halos.

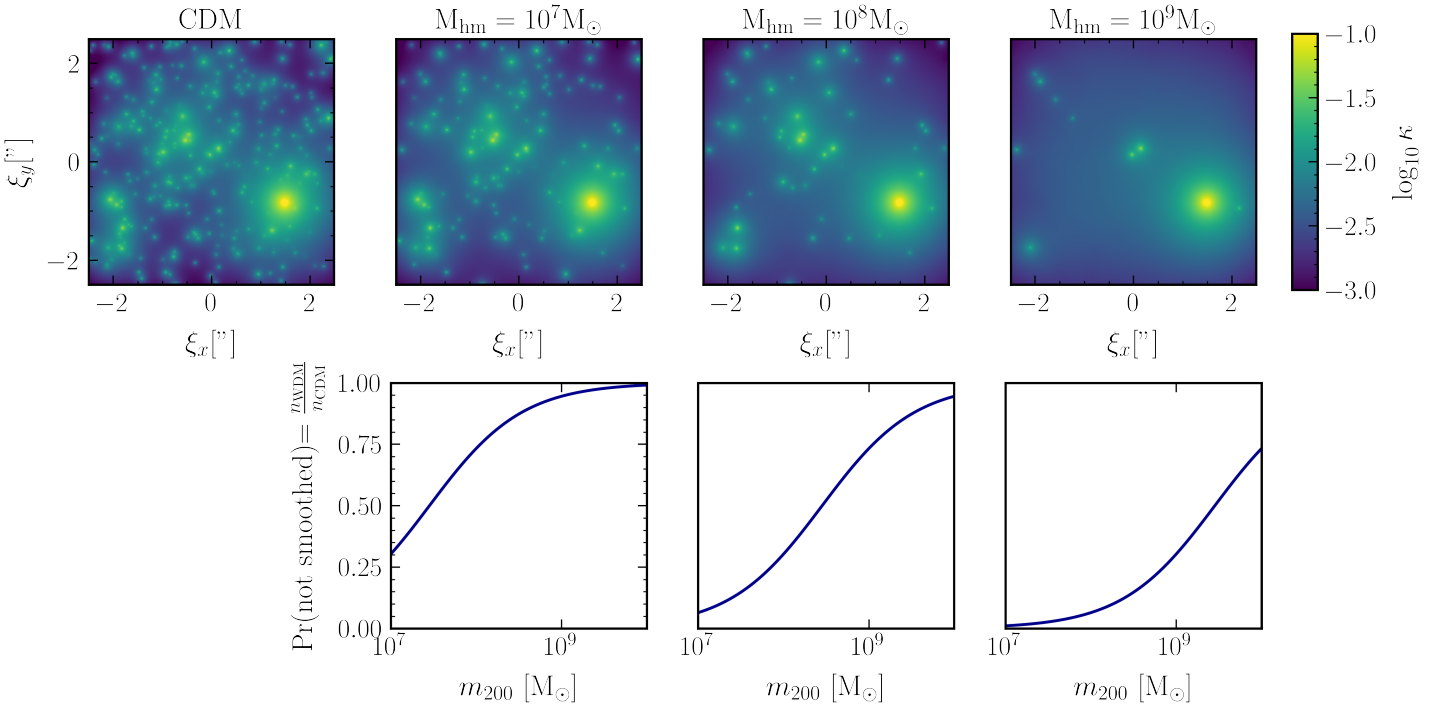

Figure 3.3: *Top*: Convergence maps for a population of LOS halos with masses sampled from the CDM HMF in the adopted mass range (Table 3.1) and projected in the lens plane following the prescription from Ref. [222]. In the second to fourth columns, we imitate the effect of WDM with different cutoff masses (as labeled in the titles) via our smoothing scheme. *Bottom*: We show the probability with which LOS halos do not get smoothed, equal to the ratio between the WDM and the CDM HMF (see Equation (3.33)). The smoothing is stochastic, so for each realization of the smoothing different halos are smoothed.

Smoothing substructures

The observational signature of WDM is, thus, the absence of small-scale structures. However, in the current parameterization, this is accompanied by the removal of the corresponding mass enclosed in them, whereas in reality the mass will still be present but will be diffused throughout the smooth main halo. This effect is manifested in a correlation between the half-mode mass and the main-halo Einstein radius: suppressing more substructures leads to an increase in the inferred Einstein radius since the total mass of the system (within the image) is tightly constrained by the size of the observed ring (or arcs).

We introduce a prescription for dealing with this degeneracy, which well captures the physical reality of structure suppression due to free streaming. Halos that should be suppressed are not present because the DM particles that should make them up are freely streaming, and their mass is therefore more diluted throughout the main halo. Therefore, rather than removing or adding substructures as a response to a changing cutoff, we still sample substructures from the CDM mass function, but we smooth the displacement field generated by halos that should be suppressed based on the aforementioned prescription by Ref. [269] to hide their lensing signature. In other words, each sampled small-scale halo has a probability equal to the ratio between the WDM and the CDM HMF (Equation (3.33)) of not being smoothed.

We then effect the smoothing by convolving the deflection field of each individual sub-/LOS halo with a radially-symmetric filter

$$f \propto 1 - \exp\left(-\left(\frac{r}{r_{\rm smooth}}\right)^{n_{\rm smooth}}\right). \tag{3.34}$$

This filtering preserves the far-field lensing signature of the halo, which is only determined by its total mass. By default, we choose the smoothing scale to be equal to the halo virial radius: $r_{\rm smooth} = r_{200}$, and the smoothing exponent $n_{\rm smooth} = 2$.

In the top row of Figure 3.3 we visualize the convergence maps in the lens plane for the same realization of LOS halos drawn from CDM distributions (panel 1), and with different cutoff masses implemented with our smoothing scheme (panels 2-4). In the bottom row, we show how we decide to smooth the lensing signature of certain halos based on the ratio between the WDM and the CDM HMF (Equation (3.33)).

3.2.6 Instrumental effects

We generate mock data with comparable quality to HST observations. All images are $5'' \times 5''$ with $0.05''$ resolution (100×100 pixels). In our simulations, we do not include a point-spread function (PSF) for simplicity, but this component cannot be disregarded in real data analysis. To account for the fact that each pixel in the image corresponds

to a finite collecting region in the sky, we generate our images at a resolution eight times higher than the target resolution and downsample. In experiments we found that neglecting this effect can have a significant impact on inference results. Lastly, we add Gaussian and uncorrelated pixel noise to our observations such that the brightest pixels are approximately 30 times the noise level (after downsampling), representative of HST data.

3.3 Methodology

Constraining the fundamental properties of DM by characterizing the population of DM halos in a strong lensing image is an extremely difficult problem since the signal we are interested in has a sub-percent level influence on images dominated by statistical noise. The problem is further complicated by the large differences between images of different lensing systems.

Our ultimate goal is to compute the marginal posterior $p(\mathrm{M_{hm}}|\boldsymbol{x})$ for a single parameter of interest $\mathrm{M_{hm}}$, the half mode mass, given an observation $\boldsymbol{x}$, for which we have the generative model

$$\begin{aligned} &p(\boldsymbol{x},\boldsymbol{\eta}_{\mathrm{src}},\boldsymbol{\eta}_{\mathrm{lens}},\vec{\boldsymbol{\eta}}_{\mathbf{p}},\mathrm{M_{hm}}) \\ &\quad = p(\boldsymbol{x}|\boldsymbol{\eta}_{\mathrm{src}},\boldsymbol{\eta}_{\mathrm{lens}},\vec{\boldsymbol{\eta}}_{\mathbf{p}},\mathrm{M_{hm}})p(\boldsymbol{\eta}_{\mathrm{src}})p(\vec{\boldsymbol{\eta}}_{\mathbf{p}}|\boldsymbol{\eta}_{\mathrm{lens}},\mathrm{M_{hm}})p(\boldsymbol{\eta}_{\mathrm{lens}})p(\mathrm{M_{hm}})\ , \end{aligned} \tag{3.35}$$

The first factor on the right-hand side is the simulator, while the other factors denote the priors on the various source, lens, and DM substructures parameters as listed in Table 3.1.

Therefore, in order to derive $p(\mathrm{M_{hm}}|\boldsymbol{x})$ we need to marginalize over all the nuisance parameters $\boldsymbol{\eta} \equiv \{\boldsymbol{\eta}_{\mathrm{lens}},\boldsymbol{\eta}_{\mathrm{src}},\vec{\boldsymbol{\eta}}_{\mathbf{p}}\}$:

$$p(\mathrm{M_{hm}}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\mathrm{M_{hm}})}{p(\boldsymbol{x})}p(\mathrm{M_{hm}}) = \frac{\int \mathrm{d}\boldsymbol{\eta}\, p(\boldsymbol{\eta})p(\boldsymbol{x}|\mathrm{M_{hm}},\boldsymbol{\eta})}{p(\boldsymbol{x})}p(\mathrm{M_{hm}}). \tag{3.36}$$

This is a very high-dimensional and multi-modal integral, even for simple analytical lens and source models, due to the large population of interchangeable substructures. Therefore, it is intractable, which renders likelihood-based inference infeasible in this case.

Instead, we approximate $p(\mathrm{M_{hm}}|\boldsymbol{x})$ using SBI with the TMNRE algorithm, presented in Section 2.2.

Formally, inferring the marginal posterior for the substructure population parameter of interest would require marginalizing over all the source, lens, and substructure realizations compatible with all possible strong lensing images. However, sampling lens and source parameters from their priors would require a very large amount of

training data and a more complex network architecture when using neural ratio estimation. This has been attempted only in Ref. [70] to infer the slope and normalization of the HMF. In order to reduce the complexity of the problem and fully exploit available information in the data with limited computational resources, we propose to target one image at a time, focusing simulations and the network training on a specific observation of interest. Thanks to TMNRE, we implement this with a truncation scheme.

TMNRE generates a sequence of likelihood-to-evidence ratio estimators on both nuisance and parameters of interest for a specific observation $\boldsymbol{x}$. In multiple inference rounds, the proposal distribution for nuisance parameters is updated and constrained, based on these ratio estimators, in order for the training data to match each round more closely the observation of interest $\boldsymbol{x}$ (this can be visually appreciated in Figure 3.7, which will be discussed in more details in Section 3.4.3).

The procedure for the truncation scheme is the following. In the first inference round, we generate training data sampling the nuisance parameters from the initial prior $p(\boldsymbol{\eta})$. Then, in each round, we constrain the proposal distribution $p_\Gamma(\boldsymbol{\eta})$ for the parameters we want to marginalize over to a region Γ where the nuisance parameters are more likely to have generated $\boldsymbol{x}$ based on the ratio estimator trained in that round. In particular, we estimate the new region Γ by very conservatively truncating the previous proposal distribution $p_\Gamma(\boldsymbol{\eta})$ in the region where the ratio estimator exceeds a predetermined threshold. We set the threshold hyperparameter to $\epsilon = 10^{-5}$, which, in case of a Gaussian posterior, corresponds to truncating at $\sim 4.78\ \sigma$ [161]. We obtain the final proposal distribution for our nuisance parameters when the region Γ does not change significantly anymore between rounds.

In this work we target with TMNRE a restricted set of the nuisance parameters $\boldsymbol{\eta}$: namely, those of the analytic smooth lens and source models, $\boldsymbol{\eta}_{\rm lens}$ and $\boldsymbol{\eta}_{\rm src}$, while leaving halo parameters, $\vec{\boldsymbol{\eta}}_{\mathbf{p}}$ unconstrained.

To summarize, thanks to TMNRE, the overall analysis strategy splits into the following steps:

1. Train an inference network to constrain the source and lens parameters, $\boldsymbol{\eta}_{\rm src}$ and $\boldsymbol{\eta}_{\rm lens}$, within ranges consistent with the observation. We then generate targeted training data based on this constrained model.

2. Train an inference network to learn the marginal likelihood-to-evidence ratio for our parameter of interest, the half mode mass $\mathrm{M_{hm}}$, on the targeted training data.

Similarly to the reasoning behind the ABC rejection algorithm, which discards sampled parameters values if the generated data is too different from the observed

data, we justify this approach by noting that parameters that do not produce observations similar to $\boldsymbol{x}$ will not contribute to the integral in Equation (3.36). Restricting the input parameters in this way immensely reduces the variability of simulated data, which allows us to use simpler network architectures and fewer training examples in the next step. As a result, the inference is now *targeted* to the specific observation at hand rather than amortized over all the possible lens/source combinations from the full prior. We would like to point out that the inference is still locally amortized in the constrained proposal distribution region, and this enables empirical test of the inference result (see Section 3.4.5).

3.4 Experiment

In this section, we show our results for hierarchical population parameters inference. First, we describe the simulated data in Section 3.4.1 and the inference network architecture in Section 3.4.2. We then show how we constrain the lens and source parameters in Section 3.4.3. Next, we show our results for the HMF cutoff mass and describe how we can combine the information from different strong lensing images in Section 3.4.4. In the same section, we show our results on the DM mass inference. Finally, we directly assess the statistical behavior of the trained neural networks in Section 3.4.5.

3.4.1 Mock data generation

For this inference task we use our lensing simulator at its full capacity, by including lens, source, substructures and different DM models that depend on the half-mode-mass $\mathrm{M_{hm}}$:

$$p(\boldsymbol{x} \mid \boldsymbol{\eta}_{\mathrm{src}}, \boldsymbol{\eta}_{\mathrm{lens}}, \vec{\boldsymbol{\eta}}_{\mathbf{p}}, \mathrm{M_{hm}}) = \mathcal{N}(\boldsymbol{x} \mid \mathrm{obs}(\boldsymbol{\eta}_{\mathrm{src}}, \boldsymbol{\eta}_{\mathrm{lens}}, \vec{\boldsymbol{\eta}}_{\mathbf{p}}, \mathrm{M_{hm}}), \sigma^2). \tag{3.37}$$

In Figure 3.4 we show a gallery of twenty mock strong-lensing images we use as target observations. These mock observations have been generated with arbitrary lens and source parameters drawn from the initial prior in Table 3.1. Their peak SNR is $\sim$30, representative of HST data.

3.4.2 Neural network

For all tasks we use the same general ratio estimator architecture, as seen in Section 2.2.4. It consists of an initial compression network $C_\Phi(\boldsymbol{x})$ that maps the 100×100 pixel images into a feature vector. This feature vector is concatenated to the parameters we want to infer (e.g. to $\boldsymbol{\eta}_{\mathrm{src}}$ and $\boldsymbol{\eta}_{\mathrm{lens}}$ for tasks where they are constrained, and

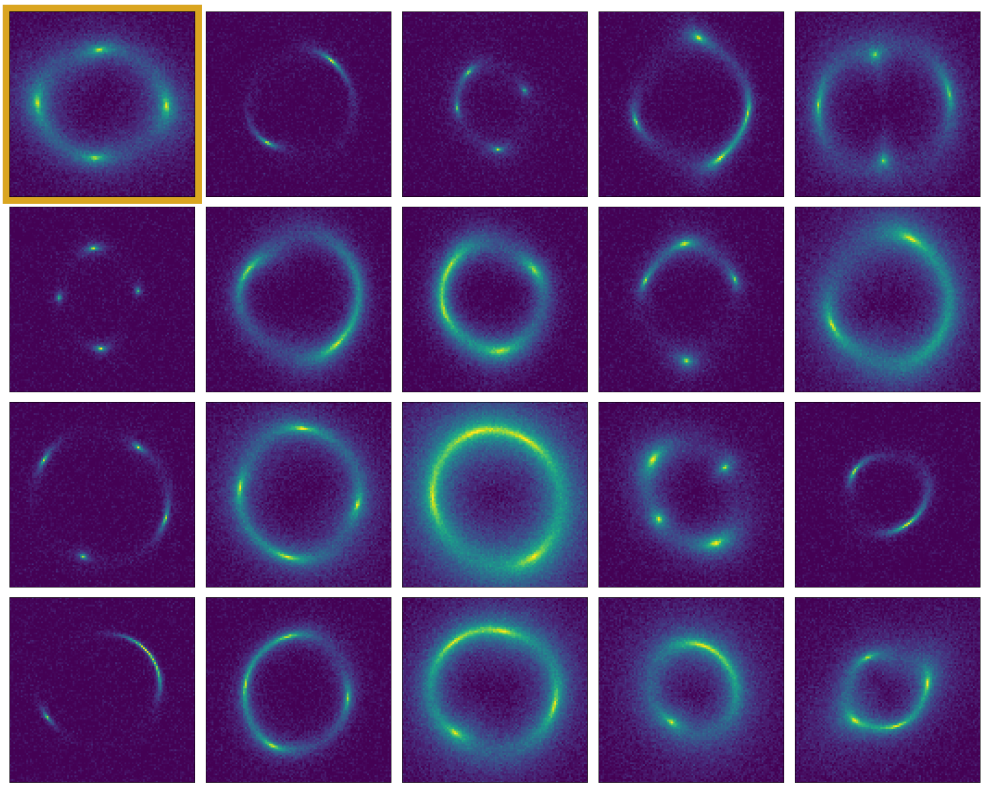

Figure 3.4: We present a gallery of twenty mock strong-lensing images we use as target observations. These mock observations have been generated with arbitrary lens and source parameters drawn from the initial prior in Table 3.1. Their peak signal-to-noise (SNR) is ∼30, representative of HST data. We analyze these images by first constraining their lens and source parameters proposal distribution in Section 3.4.3. Then, we combine them in order to infer the cutoff mass scale in Section 3.4.4. For the first one (upper left corner, framed in orange) of these images we show our results of the first part of the pipeline (Section 3.4.3) in Figure 3.6, Figure 3.7, and Figure 3.8.

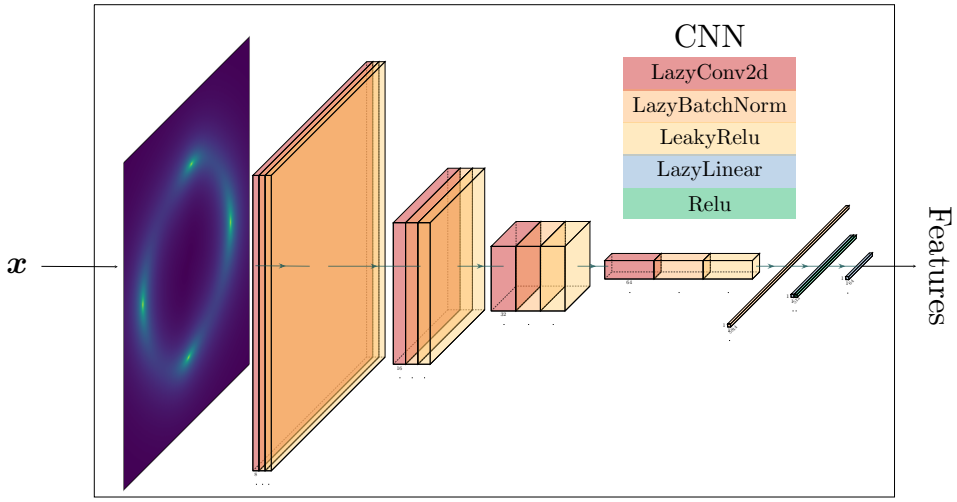

Figure 3.5: Illustration of the embedding CNN architecture used in the first part of the pipeline to constrain lens and source parameters. The observation $\boldsymbol{x}$ gets compressed into features: estimates of the best possible data summary statistic, by the CNN. In describing the CNN layers we follow `PyTorch` [109] convention. To create the illustration we have used `https://github.com/HarisIqbal88/PlotNeuralNet/`.

M_{hm}). The vector is then passed to a binary classification network $d_\Phi(\boldsymbol{x}, \boldsymbol{\theta})$ which outputs an estimate of the 1D marginal likelihood-to-evidence ratios for the parameter of interest (with separate MLPs used to estimate the 1D ratios for $\boldsymbol{\eta}_{\mathrm{src}}$ and $\boldsymbol{\eta}_{\mathrm{lens}}$). For the embedding network, in both steps of the pipeline, we adopt a simple convolutional neural network (CNN). In Figure 3.5 we show the CNN architecture used to constrain lens and source parameters. The one used to estimate the cutoff mass has a similar structure.

While we did not perform a full hyperparameter exploration, we found the batchnorm layers to be crucial for stable training of the CNN used for the macromodel ratio estimator.

3.4.3 Constraining lens and source parameters

We constrain lens and source parameters regions with TMNRE (Section 2.2) with multiple sampling and training rounds.

In total, we perform six sampling and training rounds. In each round, we simulate 10^5 observations, of which 90% are used as the training dataset, and the remaining

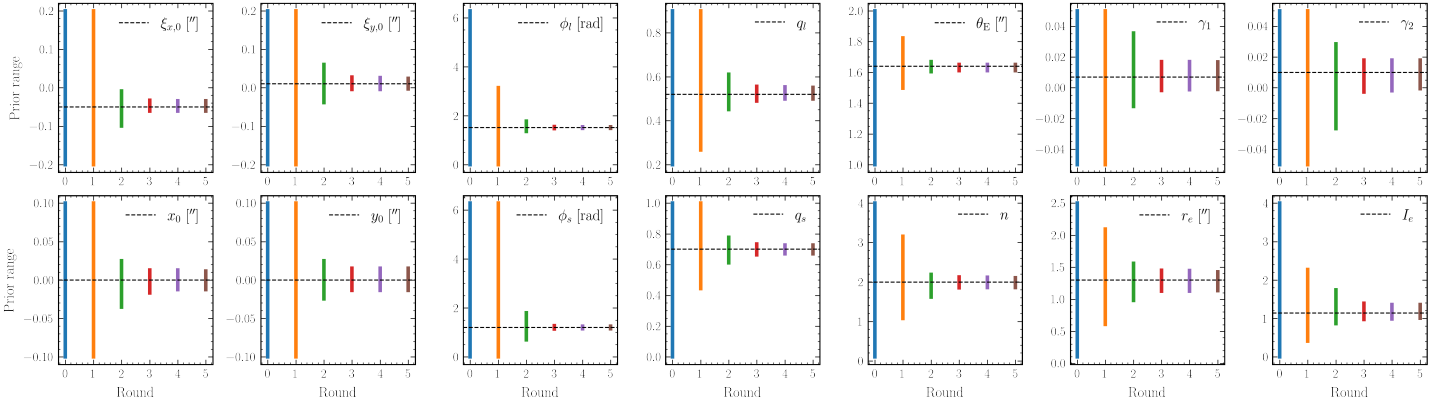

Figure 3.6: Constrained proposal distribution. Visualization of the sequential truncation of the lens and source proposal distributions over the six rounds of training. The particular target is the first mock image (framed in orange in Figure 3.4), whose parameters are depicted as black dashed horizontal lines.

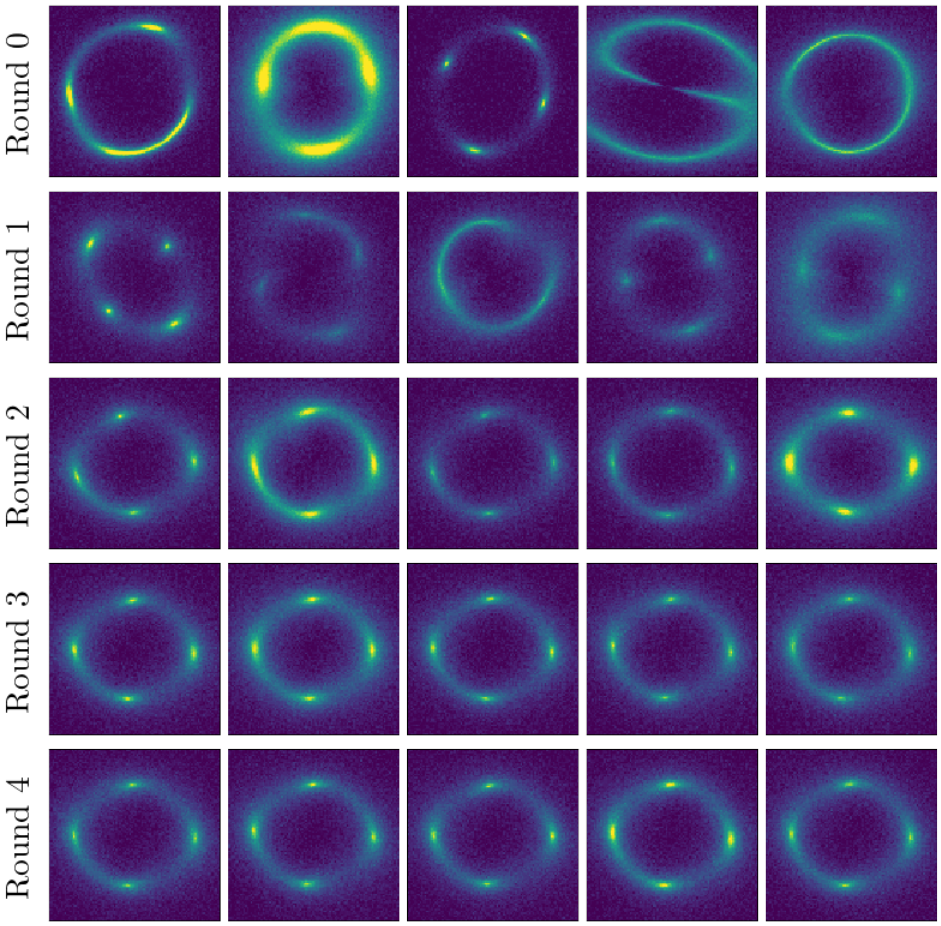

Figure 3.7: Training data targeting the first mock observation (framed in orange in Figure 3.4). In each row, we show five examples of training data for the first five rounds. In the first round, we sample our data from the initial prior shown in Table 3.1. For the following rounds, the lens and source parameters are sampled from the constrained proposal distributions, obtained by evaluating the network trained with the previous round dataset on our target observation (see Section 3.4.3). It is evident that with each round the training data more closely resembles the target image $\boldsymbol{x}_0$.

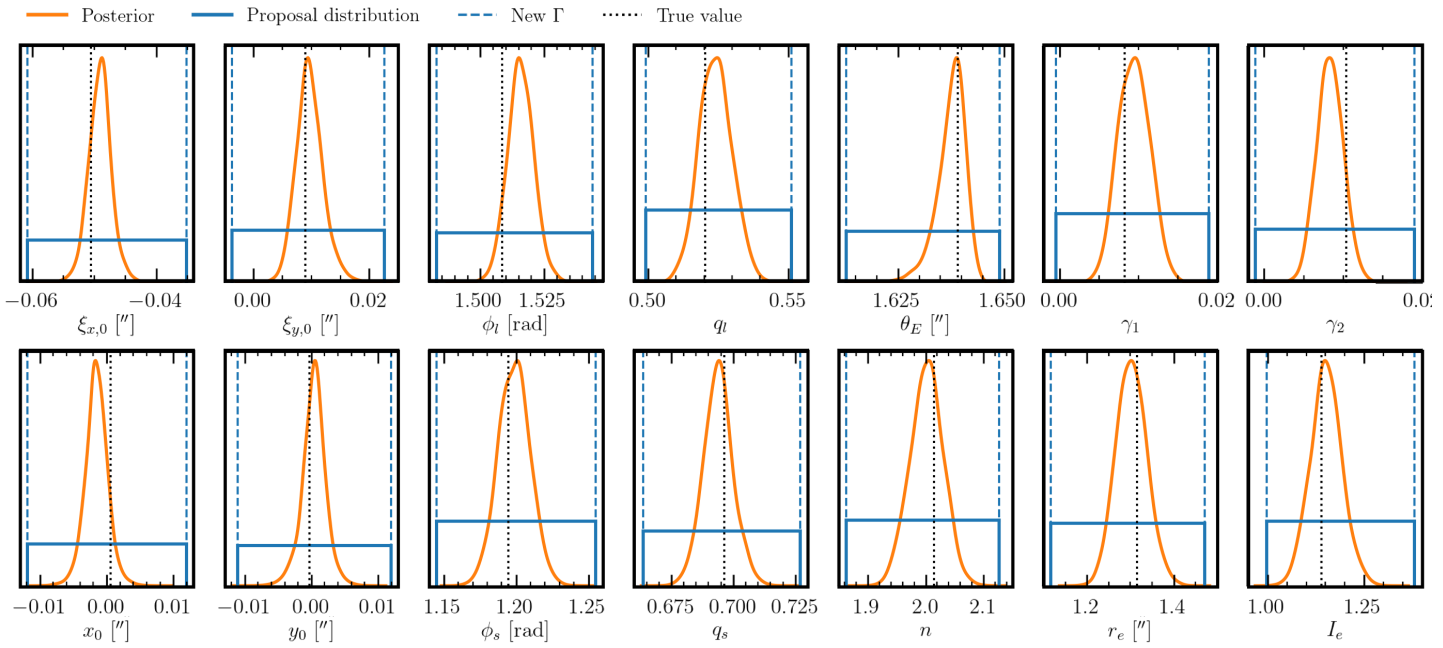

Figure 3.8: Lens and source parameters posteriors. In solid blue we show the last round of constrained proposal distributions for the first (upper left corner, framed in orange) target image in Figure 3.4. The dotted black lines correspond to the true lens and source parameters values with which we have generated the target image. In orange, we show the estimated posteriors for lens and source parameters in the last training round. Based on the predetermined threshold, the new bounding limits Γ (dashed blue) do not change significantly from the previous constrained proposal distribution region, so it is not possible to constrain the proposal distribution more and we stop the truncation procedure.

10% as the validation dataset. Evaluations of the network on the mock target image are used to truncate the training data proposal distribution after each round, so that the region for lens and source parameters is targeted. The first training round is performed on the dataset generated from the initial source and lens parameters priors, shown in Table 3.1. In Figure 3.6 we show the initial prior and the following constrained proposal distributions. It can be seen that after the first round just a few of the parameters proposal distributions get truncated, e.g. the Einstein radius. By having truncated these initial parameters, in the following rounds the other parameters can be better learned by the network and so constrained. In Figure 3.7, we show samples from the first five training datasets, which demonstrate that the constrained regions are indeed the ones that are likely to produce data similar to the targeted image $\boldsymbol{x}$. After the sixth round of training, it is not possible anymore to truncate the proposal distribution region based on the predetermined threshold, as seen in Figure 3.8. The truncation scheme has then efficiently identified the constrained region for lens and source parameters consistent with the targeted observation. Using the last constrained dataset, it is then easier in the second step of the pipeline to train a marginal neural ratio estimator to perform the final inference on the cutoff mass.

We would like to stress that these constrained proposal distributions correctly account for lens and source parameters uncertainties. In all our simulated data, the substructure parameters $\vec{\boldsymbol{\eta}}_{\mathbf{p}}$ are randomly sampled from their prior, in order to account for the presence of substructures. This has the desirable outcome of approximately accounting for the average effect that an additional mass component has on the main lens parameters (e.g. inferring an unbiased Einstein radius) and contributes to the source and lens uncertainties.

3.4.4 Dark matter inference

For the second step of the pipeline, we train an inference network to learn the cutoff mass on the last constrained dataset.

From initial tests, we have found that features from a single image are very hard to learn for the classifier, resulting in a very noisy ratio estimator. In order to reduce the estimator uncertainty, we then train the cutoff mass classifier on a dataset $X^N = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ of N different observations. For each observation, first, we constrain its lens and source parameters as explained in Section 3.4.3. Then, we train the cutoff classifier on the concatenation of the features coming from their embedding networks, effectively learning $r(\mathrm{M_{hm}}; X^N)$. Note that the images in one dataset are sampled with the same cutoff mass $\mathrm{M_{hm}}$, but different lens, source, and substructures realizations. In fact, our final goal is to apply the full pipeline to real data, which will all have different source, lens, and substructures configurations, but will have encoded the same DM properties.

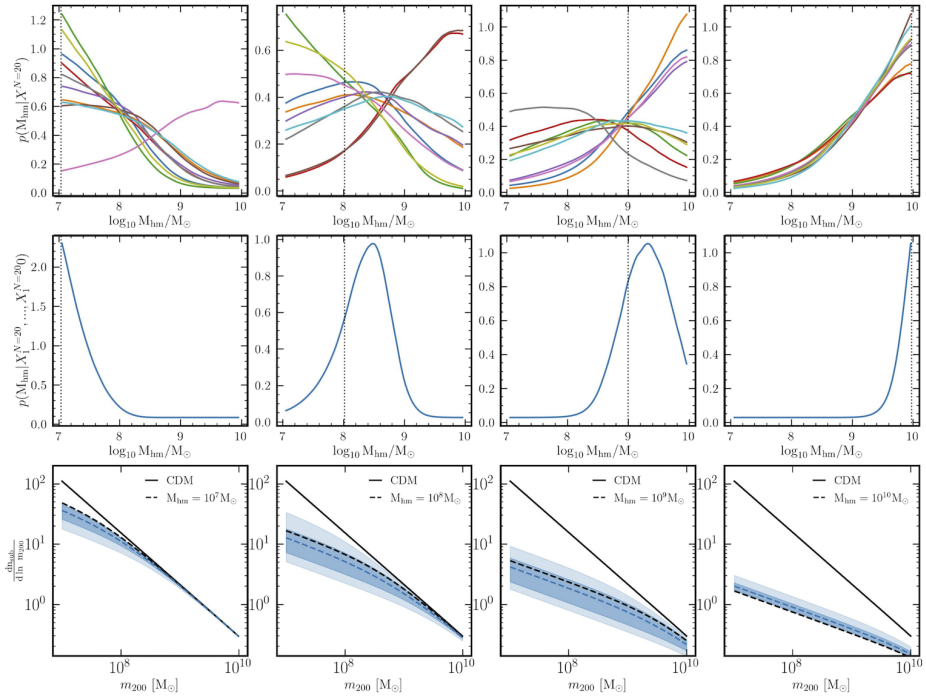

Figure 3.9: *Top:* Approximate posteriors for the half-mode mass derived from 10 different sets of 20 images. The dotted black line represents the true value of the half mode mass with which we have generated the images ($10^7, 10^8, 10^9, 10^{10}$ $M_\odot$). *Middle:* We show the approximate posterior resulting from the combination of the $M = 10$ different posteriors shown in the first column, as explained in the text (Section 3.4.4). *Bottom:* Subhalo mass function constraints derived from the cutoff mass posterior shown in the second column. The black solid line shows the CDM subhalo mass function according to Equation (3.22), whereas the black dashed one shows the WDM subhalo mass function according to Equation (3.33), given the true cutoff mass shown in the label. The blue dashed line shows the mean of the WDM subhalo mass function obtained by sampling 1000 samples from the cutoff mass posterior shown in the second panel and using this value in Equation (3.33). We also show the central 68 and 95 percentiles as shaded bands. These plots show how uncertain the subhalo mass function is under the assumption that it has the functional form in Equation (3.33) with parameters from Ref. [269].

In the first row of Figure 3.9 we show the results from the inference network on ten test sets of lenses generated with a $\mathrm{M_{hm}}$ value of $10^7, 10^8, 10^9$, and 10^{10} $\mathrm{M_\odot}$. Each curve is the posterior obtained for a set of $N = 20$ lenses. Each of the mock observations has lens and source parameters sampled from their own final constrained proposal distribution, and different substructure population.

Now that we have reduced the estimator noise, it is straightforward to perform inference on a group of sets of images by combining their ratios. Given a dataset $X^N = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ of images, the combined ratio for multiple M datasets is simply given by $r(\mathrm{M_{hm}}; X^N_M,) \propto \prod_{i=1}^{M} r(\mathrm{M_{hm}}; X^N_i)$, where the proportionality is a ratio of evidences, independent of the parameter value, so it only accounts for a proper normalisation Ref. [70, 153]. In the second row of Figure 3.9 we show the results for the combination of the $M = 10$ different posteriors shown in the first column.

In the third row we show a combined posterior for the WDM mass function from 200 images ($M = 10$ sets of $N = 20$ images). These plots show the uncertainty in the subhalo mass function under the assumption that it has the functional form in Equation (3.33) with parameters from Ref. [269].

These first results show that our method is sensitive to the low-mass end of the HMF, and that we have unbiased results from combining just 10 sets of 20 observations, given that in the second panel of Figure 3.9 the true input value for the half-mode mass $\mathrm{M_{hm}}$ is consistently contained within the estimated posterior. In Section 3.4.5 we will show a more sophisticated method to assess the statistical behavior of our inference results.

Furthermore, we can translate the constraints we obtain on the cutoff mass to constraints on the WDM mass given the mapping between those two quantities defined in Section 3.2.5. In Figure 3.10 we show our results for the WDM mass. Each column corresponds to a different cutoff mass input value, so a different WDM mass. In the first row, we plot five examples of the combined posterior density for $\log_{10} \mathrm{M_{hm}}$ of $M = 10$ sets of $N = 20$ observations. In the second row, we show the corresponding color-coded five examples for m_{WDM}. In this case, we just transform the posterior from the first row using the parameterization shown in Section 3.2.5, so we assume a flat prior on $\log_{10} \mathrm{M_{hm}}$. Finally, in the last row, we show the WDM mass posterior densities assuming a flat prior on the latter. The posteriors in the second and third row are not actually the same because a flat prior $\log_{10} \mathrm{M_{hm}}$ is different from a flat prior on m_{WDM}.

3.4.5 Credible interval testing

We would like to directly test and validate the statistical behavior of our inference results by determining the expected coverage of the ratio estimator produced by the

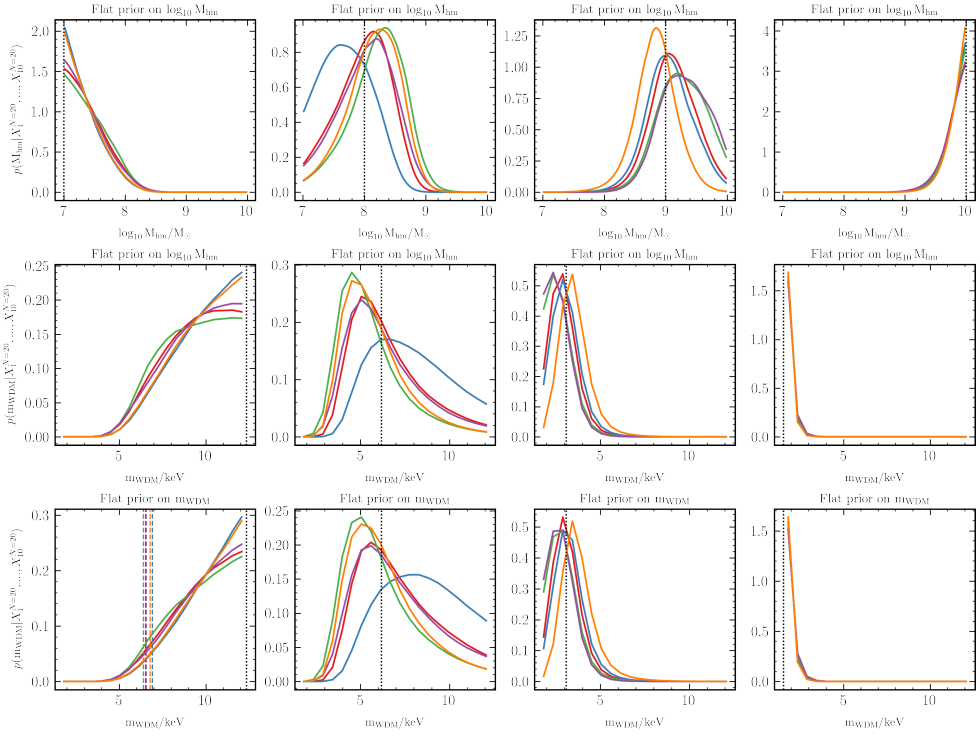

Figure 3.10: *Top:* We show five examples of combined posterior of $M = 10$ sets of $N = 20$ observations in terms of the cutoff mass (as the second row in Figure 3.9). The dotted black line represents the true input value of the half mode mass with which we have generated the analyzed mock observations ($10^7, 10^8, 10^9, 10^{10}$ $M_\odot$). *Middle:* Same results as shown in the first column but for the WDM mass. The dotted black line represents the true value of the WDM mass with which we have generated the analyzed mock observations, given the mapping between DM cutoff and DM mass in Section 3.2.5. The WDM mass posteriors assume a flat prior on the cutoff mass. *Bottom:* Same results as shown in the first column but for the WDM mass and assuming a flat prior on the latter. In the first plot of the row, we show for the five examples the expected 95% credible lower limit on the WDM mass for the highest value of our prior distribution.

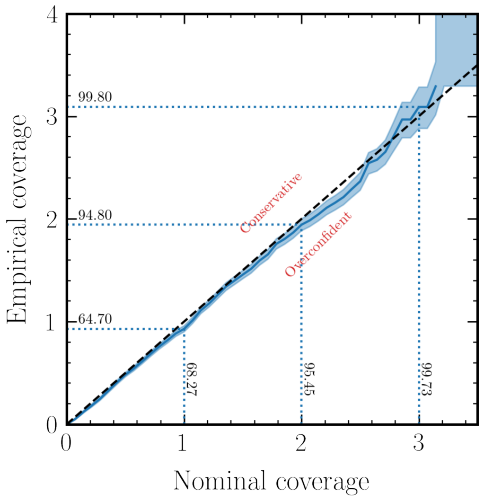

Figure 3.11: Empirical versus nominal expected coverage probabilities for the cutoff mass inference network. In case the line lies above (below) the black dashed diagonal line, the credible intervals are conservative (overconfident) and contain the true value with a frequency higher (lower) than nominally expected. We show the empirical (nominal) probabilities as horizontal (vertical) text.

network, as described in Section 2.3. The goal is to compare the nominal and empirical expected coverage probabilities of estimated Bayesian credible intervals, which should coincide for a well-calibrated estimator. For the statistical formalism and definition of credible region and expected coverage probability, we refer the reader to Ref. [171]. In brief, an ideal estimator has matching empirical and nominal expected coverage, a conservative one predicts lower credibility than empirically obtained, and an overconfident one has higher nominal than empirical credibility. In plots like Figure 3.11, the line for an ideal ratio estimator should perfectly align with the diagonal, whereas for a conservative (overconfident) estimator, it will lie above (below) the diagonal. In combination with visually checking the posteriors, this test supports the accuracy of the posterior estimator and is also particularly useful when one does not have access to the ground truth against which to compare the results. In Figure 3.11 we show the empirical versus nominal expected coverage probabilities for the cutoff mass inference network. We can see that the inference network for the half-mode mass has converged with good expected coverage.

3.5 Discussion

In this section, we discuss the improvements to the model and inference question which need to be addressed before we can safely apply our pipeline to the analysis of real data.

First, we have neglected effects such as inadequate *lens light* subtraction and assumed the lens light to be known (this will be accounted for in Chapter 4). Regarding the *noise* model, we did not account for correlated pixel noise due to instrumental effects including the telescope's PSF, multi-band observations, drizzling (e.g. see Ref. [68]), and even complex noise with an unknown likelihood function.

In this work, we have employed an analytic parameterization (the Sérsic profile) as a lensed source light distribution model, which is adequate to analyze low-resolution images. However, to accurately model higher-fidelity lensing observations, such as those from on-going (e.g. HST) and future (e.g. JWST, ELT, SKA) telescopes, more *complex source models* need to be employed. Existing models, in order of complexity, are regularized pixellation of the source plane (see, e.g., Refs. [41, 119, 270]), source modeling through basis functions (e.g. shapelets [271] or wavelets [272]) attached to the source plane, and deep learning approaches (see, e.g., Refs. [230, 273]). The ability to accurately and precisely reconstruct the complex morphology of strong-lensing sources is of the utmost importance, as to disentangle the source surface brightness inhomogeneities from the percent-level fluctuations introduced by substructures in the lens. We anticipate that using sources with more complex morphologies will result in higher sensitivity to the DM cutoff mass, provided that it is possible to model these sources. In fact, the residuals between the image of an extended source lensed by the total lens potential (accounting for substructures), and that of the same source lensed only by the main lens component are proportional to the gradient of that source evaluated in the image plane [224, Equation 16].

Regarding WDM modeling, validation of our smoothing scheme (Equation 3.2.5) is required to accurately account for DM free-streaming effects. Moreover, we should account for uncertainties due to the assumed halo density profile by considering different DM distributions around galaxies (see, e.g., Ref. [274] for a review). We note that, thanks to its flexibility, our pipeline can incorporate any arbitrary DM model, as long as it specifies the form of the HMF and the density profiles of individual substructures.

Finally, we would like to draw the reader's attention on the fact that in our modeling we assume that the halo mass of the lens is known exactly from its Einstein radius (see Equation 3.2.4). This is a strong assumption that has as a consequence the separation of substructure parameters $\vec{\boldsymbol{\eta}}_{\mathbf{p}}$ and lens parameter $\boldsymbol{\eta}_{\rm lens}$ once we marginalize the posterior probability over the halo mass in Equation (3.35). In particular, the

second inference question we have addressed in this work in Section 3.4.4, constraining the cutoff mass of the subhalo mass distribution, is then a simplified version of the real one, which is to simultaneously determine the halo mass and subhalo mass distribution of the lenses from real data (see, e.g., Ref. [226]).

While this work used simple mock lenses, TMNRE makes it possible to add realism and parameters to a simulator without significantly altering the inference procedure, or necessarily increasing the simulation budget [78]. It should therefore be straightforward to incorporate these various complexities we ignored in this work without fundamentally modifying the inference pipeline.

3.6 Conclusions

Measuring the collective properties of DM halos on sub-galactic scales by means of their gravitational effect is an important probe of the fundamental nature of DM. One of such probes, strong gravitational lensing, has sparked much interest over the last few years. Moreover, the development of fast and accurate techniques to extract information from strong lensing images is well motivated by the wealth of new high-resolution strong lensing observations that will become available in the near future.

In this work, we have presented the first step towards a new neural simulation based inference pipeline (Chapter 2) to analyze present and future strong gravitational lensing systems in order to constrain the cutoff in the DM HMF, and so the DM mass (Section 3.4). To this end, we have used a recent machine learning development, TMNRE, that makes it possible to *target* the analysis to a specific observation rather than amortize over all possible variations in lensing systems, making inference more efficient and precise. Thanks to TMNRE, we overcome the computational challenges of traditional MCMC, nested sampling and trans-dimensional MCMC methods, by directly learning the marginal posterior for the parameter of scientific-interest from the observation. Moreover, TMNRE leverages neural networks to directly learn the best summary statistic possible from the full input data, without having to compress the observation into hand-crafted summary statistics, like for ABC frameworks. The method is applicable to simulators with unknown likelihood functions and large or even variable numbers of input parameters. Lastly, the resulting inference networks can be poked and prodded to confirm they are statistically well-behaved. This work is then a step forward towards making the analysis of strong lensing images for DM science faster, more efficient, and more accurate.

Our key results can be summarized as follows:

TMNRE enables direct marginal inference. Thanks to marginalized inference,

the TMNRE based analysis is able to correctly propagate the lens and source parameters uncertainties, and account for the presence of a population of substructures, when estimating the marginal posterior of interest.

TMNRE enables direct targeted inference. Thanks to our targeted approach, we are able to correctly estimate the lens and source parameters uncertainties, and use the final lens and source parameters truncated proposal distributions (see Section 3.4.3) to generate a targeted training dataset in order to infer the DM cutoff.

TMNRE enables statistical checks. Since the inference networks learned by TMNRE are locally amortized over a range of potential observations, we were able to test their statistical consistency. Our checks confirm that TMNRE on average produces posteriors with the correct width (see e.g. Figure 3.11). Such tests would be extremely expensive with likelihood-based inference since they would require rerunning the sampling machinery on numerous mock observations.

TMNRE enables hierarchical inference. We demonstrated that our framework is able to statistically extract the DM cutoff mass signal from a population of small-scale dark matter halos in the $[10^7,\ 10^{10}]$ $M_\odot$ mass range, by performing hierarchical inference on up to 200 observations (Section 3.4.4 and Figure 3.9). A cutoff mass posterior translates into a posterior on the WDM mass, given the mapping in Section 3.2.5. What we find is an expected 95% credible lower limits around 6.5 keV in the case of the scenario closest to CDM (see the bottom left panel in Figure 3.10), given the adopted prior and the various assumptions discussed in Section 3.5.

In this work, we have demonstrated that, in principle, the DM cutoff mass signal can be statistically extracted from a population of small-scale dark matter halos by a neural network using TMNRE. A part from the possible framework improvements discussed in Section 3.5, an interesting direction for further work is the use of TMNRE for model comparison. While here our ratio estimators were trained to compute the likelihood-to-evidence ratio, as pointed out in Ref. [153] it is possible to learn other ratios of densities. In particular ratio estimators can be used to learn the Bayes factor for assessing the strength of the evidence for different models. This could be used to determine whether an image contains a perturber or not, and to map the minimum-detectable perturber mass as a function of its position.

Overall, we believe using TMNRE to measure dark perturbers' population parameters as described in this work provides a promising path towards uncovering the identity of dark matter.

4 Scalable simulation-based inference

In recent years, there has been a remarkable development of SBI algorithms, and they have now been applied across a wide range of astrophysical and cosmological analyses. There are a number of key advantages to these methods, centered around the ability to perform scalable statistical inference without an explicit likelihood. In this chapter, we propose two technical building blocks to a specific sequential SBI algorithm, TMNRE. In particular, first we develop autoregressive ratio estimation with the aim to robustly estimate correlated high-dimensional posteriors. Secondly, we propose a slice-based nested sampling algorithm to efficiently draw both posterior samples and constrained prior samples from ratio estimators, the latter being instrumental for sequential inference. To validate our implementation, we carry out inference tasks on three concrete examples: a toy model of a multi-dimensional Gaussian, the analysis of a stellar stream mock observation, and finally, a proof-of-concept application to substructure searches in strong gravitational lensing.

This chapter is based on work from [2].

4.1 Introduction

Our understanding of the universe has the potential to be revolutionized by the exponentially growing influx of high quality data from current and upcoming astrophysical or cosmological surveys [275]. Data from current and near-future facilities (e.g. Euclid [42], JWST [12], Rubin-LSST [10], ELT [276], Gaia [277], SKA [278], CTA [279]) will exceed the peta- and exa-byte threshold. Given this status, it is crucial for the community to develop innovative data analysis pipelines, statistical algorithms, data compression techniques, and search pipelines. These will need to handle not only the escalating amount of data, but also its increasing resolution, which directly translates into growing model complexity.

In general, to obtain information about physical models from the data, one has to solve the "inverse problem", i.e. given an observation, the goal is to infer the parameters of a specific model (or set of models) that are most likely to have generated it. The main tools to solve inverse problems for modern astrophysical and cosmological data analysis have been sampling-based inference methods like MCMC [93, 94] and nested sampling [95, 96, 100] techniques. However, these methods often rely on approximate likelihoods, and the time needed to reach convergence scales poorly with the dimensionality of the explored parameter space. More modern methods are taking up this latter challenge, including gradient-based algorithms such as Hamiltonian Monte-Carlo [280], or slice-sampling techniques [99, 100].

Novel techniques in the field of SBI are starting to overcome these challenges (for a recent review see Ref. [52]), and have recently gained significant popularity due to a number of appealing features (see e.g. Refs. [5, 55, 59, 64, 67, 70, 78, 81, 90, 281–284] for examples of method development and applications across cosmology and astroparticle physics). First of all, SBI methods do not require an explicit model of the data likelihood, but instead access its information implicitly via a stochastic simulator, which maps input parameters to data. Secondly, SBI techniques are able to directly target marginal posteriors for the parameters of interest [55, 165], which typically improves scalability with the dimensionality of the parameter space. This is because an arbitrarily large number of nuisance parameters can be included whilst targeting the same set of parameters of interest.[1] Lastly, sequential SBI approaches that continuously use the acquired inference knowledge to guide the simulator into the relevant part of the parameter space through active learning, have been shown to be particularly simulation efficient [124].

Whilst there are a wide range of SBI algorithms (we refer the reader to Section 2.1 for a broad overview), the main focus of this work will be TMNRE [120, 160]. This is a sequential implementation of the general neural ratio estimation technique [153] that composes well with marginalisation. In essence, neural ratio estimation trains a neural network to approximate the posterior-to-prior ratio by solving a binary classification problem (for more details see Section 2.2.1). TMNRE has been shown to be successful in a number of different physics applications, such as cosmic microwave background analyses [78], strong lensing image analysis [1, 5], supernovae Ia cosmology [81], gravitational wave parameter inference [64, 65], along with other applications [4, 80, 90, 285].

So far, TMNRE applications have focused on estimating low-dimensional ($d \lesssim 2$) marginal posteriors for the parameters of interest. This choice was motivated by the

[1] It is important to note that if all parameters contribute equally to the data variance, the implicit data distribution will become noise-dominated. Thus, when referring to scaling to arbitrary number of variables, the data variance is implicitly kept fixed. This limit remains a challenge for sampling-based methods, but is tractable in the SBI framework.

fact that directly targeting these low-dimensional marginal posteriors alleviates sampling problems associated with high dimensionality, whilst still maintaining much of the information relevant for scientific conclusions. Nevertheless, there are situations when focusing on low-dimensional marginal posterior estimates makes TMNRE inherently inefficient, for example when the marginals of interest are highly correlated (or multi-modal). In these cases it is therefore necessary to produce accurate high-dimensional joint estimates to account for these correlations, since higher-dimensional structure can be obscured by low-dimensional projections. This limitation becomes especially relevant for the sequential aspect of the algorithm, which cannot leverage the information regarding the correlations while guiding the simulations. However, the necessity for high-dimensional joint estimates clashes with one know failure mode of density ratio estimators, the so called "density-chasm problem", described in Ref. [286]. In essence, density ratio estimators can fail whenever the gap (in the sense of Ref. [286]) between the two densities is large, since the binary classifier can obtain almost perfect accuracy with a relatively poor estimate of the density ratio. This problem is exacerbated in high-dimensions.

The present work is a step forward towards addressing these competing limitations. Here, we propose two new building blocks of the TMNRE framework: an *autoregressive* implementation of neural ratio estimation for scalable, high-dimensional (marginal) posterior inference, and a *slice-based nested sampling algorithm* to efficiently draw not only posterior samples, but also constrained prior samples. The latter set of samples, as was discussed in Section 2.2.3, are necessary for TMNRE's implementation of active learning.

In recent years, autoregressive models have shown great potential in scaling to high-dimensional distribution estimation problems, see e.g. Refs. [287–289]. The term "autoregressive" originates from the time-series modeling literature, where it refers to the practice of using previous time-step observations to predict the value at the current time step. Autoregressive models function in a similar fashion by decomposing a d-dimensional joint density into a product of d 1-dimensional conditional distributions, as in Equation (4.1). An autoregressive model is then defined by the parameterization of all d conditionals.

We introduce the slice sampling component because neural ratio estimators do not come with sampling functionalities. In general, to sample from the estimated (marginal) posterior, we require Monte Carlo sampling algorithms, especially in high dimensions. Moreover, we are interested in obtaining constrained prior samples for the purposes of active learning. In this context, whilst nested sampling was primarily born as a general purpose integration algorithm in high-dimensions, it has primarily been applied to perform Bayesian inference, in particular to estimate the Bayesian evidence and the posterior. Interestingly, at the heart of nested sampling algorithms resides also the problem of constrained prior sampling [97]. Motivated by the success

of slice-based nested samplers [99, 100], we implement a custom slice sampler to draw *both* posterior and constrained prior samples.

The structure of this chapter is as follows. We introduce this work's main contribution: autoregressive neural ratio estimation and prior truncation through slice-based nested sampling, in Section 4.2. In Section 4.3, we present results on a toy example, and applications to stellar streams and substructure parameter inference in strong gravitational lensing images, which initially motivated the development of the presented tools. Finally, we provide some discussion regarding possible limitations and outlook, concluding in Section 4.4.

Code. We release a publicly available implementation of the autoregressive ratio estimator model in the TMNRE package `swyft`[2], and an implementation of the slice-based nested sampler in `torchns`[3].

4.2 Methodology

Up to now, applications of TMNRE have typically focused on estimating low-dimensional ($d \lesssim 2$) marginal posteriors. On the other hand, whilst correlations between two parameters can always be estimated by training an appropriate ratio estimator, doing this for all pairwise combinations of a large number of parameters becomes quickly infeasible. Furthermore, even if it were done, two-dimensional correlations do not provide any information about higher order correlations since they are just projections. In certain scenarios, this might be crucial for calibrating the quality and accuracy of the inference methods and strongly motivates the development of techniques able to robustly estimate higher-dimensional posterior distributions.

As discussed in Section 4.1, however, there are challenges related to both the estimation of high-dimensional ratios and subsequently sampling from them. Here, we propose two new building blocks of the TMNRE framework to address these issues: *autoregressive NRE* for scalable, multi-dimensional (marginal) posterior estimation, and *slice sampling* to efficiently sample from a multi-dimensional posterior and truncated prior.[4]

[2] https://github.com/undark-lab/swyft
[3] https://github.com/undark-lab/torchns
[4] This work has been informed by private communications surrounding a complementary paper in preparation [290].

4.2.1 Autoregressive Neural Ratio Estimation

Across the various SBI techniques, autoregressive models have been developed in the context of density estimation algorithms, NPE [288, 289] and NLE [149], and have been shown to be among the best performing density estimators [124]. They have been successfully applied to the Dark Energy Survey and global 21cm signal experiments [291], and for deep generative models for galaxy image simulations [292].

In a nutshell, autoregressive models turn the estimation of a d-dimensional joint density into the estimation of d 1-dimensional conditional densities using the chain rule of probability

$$p(\boldsymbol{\theta} \mid \boldsymbol{x}) = p(\theta_1 \mid \boldsymbol{x}) \prod_{i=2}^{d} p(\theta_i \mid \boldsymbol{x}, \boldsymbol{\theta}_{1:i-1}) \,, \tag{4.1}$$

where we have introduced the compact notation $\boldsymbol{\theta}_{1:i-1} \equiv \{\theta_1, ..., \theta_{i-1}\}$. One can thus define an autoregressive model simply by specifying a parameterization of all d conditionals.

Within the NRE framework, it is also possible to perform *conditional inference*. In this case, the binary classifier must be informed about the value of the parameters we want to condition over. More specifically, in order to learn the conditional posterior for parameter θ_i given parameter θ_j (with $j \neq i$), the binary classifier will be shown as positive training example $\boldsymbol{x}, \theta_j, \theta_i \sim p(\boldsymbol{x}, \theta_j, \theta_i)$ and as negative training examples $\boldsymbol{x}, \theta_j, \theta_i \sim p(\boldsymbol{x}, \theta_j)p(\theta_i)$. We can then express the conditional ratio trained on $(\{\boldsymbol{x}, \theta_j\}, \theta_i)$ pairs as

$$r(\theta_i; \boldsymbol{x}, \theta_j) \equiv \frac{p(\theta_i \mid \boldsymbol{x}, \theta_j)}{p(\theta_i)} \,. \tag{4.2}$$

It is possible to condition more than one parameter θ_i on more than one variable θ_j just by correctly passing to the classifier positive and negative training examples as defined above.

We can then use 1-dimensional versions of these ratio estimators, conditioned on multiple variables, to implement autoregressive NRE. More in details, one way to estimate our ratio of interest $r(\boldsymbol{\theta}; \boldsymbol{x})$, as defined in Equation (2.20), auto-regressively is by considering the following components. The first quantity, A, estimates the ratio between the joint posterior distribution and the independent marginal priors,

$$\begin{aligned} A &= \frac{p(\boldsymbol{\theta} \mid \boldsymbol{x})}{\prod_{i=1}^{d} p(\theta_i)} \\ &= \frac{p(\theta_1 \mid \boldsymbol{x})}{p(\theta_1)} \prod_{i=2}^{d} \frac{p(\theta_i \mid \boldsymbol{x}, \boldsymbol{\theta}_{1:i-1})}{p(\theta_i)} \\ &= r(\theta_1; \boldsymbol{x}) \prod_{i=2}^{d} r(\theta_i; \boldsymbol{x}, \boldsymbol{\theta}_{1:i-1}) \,. \end{aligned} \tag{4.3}$$

The second component, B, models the dependencies between the model parameters,

$$\begin{aligned} B &= \frac{p(\boldsymbol{\theta})}{\prod_{i=1}^{d} p(\theta_i)} \\ &= \prod_{i=2}^{d} \frac{p(\theta_i \mid \boldsymbol{\theta}_{1:i-1})}{p(\theta_i)} \\ &= \prod_{i=2}^{d} r(\theta_i; \boldsymbol{\theta}_{1:i-1}) \ . \end{aligned} \tag{4.4}$$

Our key quantity of interest is then obtained as

$$r(\boldsymbol{\theta}; \boldsymbol{x}) = \frac{p(\boldsymbol{\theta} \mid \boldsymbol{x})}{p(\boldsymbol{\theta})} = A/B \ . \tag{4.5}$$

It is important to highlight the role of the B component in this algorithm. In the case of sequential inference, even if the initial prior distributions from which parameters are drawn are independent, the sequential proposal distribution will account for non-trivial correlations between parameters in the constrained prior region (via the condition in Equation (2.27)). It is therefore crucial to properly account for these correlations between parameters through B since they will implicitly be present in the training data.

Also note that in both the definitions of A and B, we have used the notation introduced in Equation (4.2) for conditional ratios. An alternative formulation of an autoregressive model for ratio estimation that uses a single network to model both components is presented in Appendix 4.A.

The key advantage of autoregressive NRE lies in its ability to handle intricate dependencies among variables by focusing on one parameter at a time. In "vanilla" NRE, for high-dimensional parameter spaces, the discriminating power of the binary classifier can be quickly saturated leading to a poor estimate of the joint density ratio. Here, the full joint distribution is modeled through 1-dimensional conditional ratios, leading to a more stable and accurate estimation of the full joint ratio. However, the presented autoregressive NRE method does inherit the generic drawbacks of autoregressive models. Perhaps the most relevant is their sensitivity to the order of the conditional probabilities, since in practice it is difficult to know which of the factorially many orders is the most efficient in each case [289]. A possible solution was presented in Ref. [293], that introduced an efficient procedure to simultaneously train an autoregressive model for all possible orderings of the variables. We will study this effect explicitly in the stellar streams example given in Section 4.3.2.

4.2.2 Prior truncation strategies

The original prior truncation scheme proposed in the TMNRE formalism [160] is a parameter-wise truncation based on 1-dimensional marginal ratio estimators. As a result, the prior gets restricted to a truncation region that has the shape of a hyper-rectangular box ("box" truncation),

$$p_{\Gamma}^{(R)}(\boldsymbol{\theta}) = \frac{1}{Z}\mathbb{I}(\theta_1 \in \Gamma_{\theta_1}^{(R-1)}) \times \cdots \times \mathbb{I}(\theta_d \in \Gamma_{\theta_d}^{(R-1)}) p(\boldsymbol{\theta})\,, \tag{4.6}$$

where we have introduced the indicator function $\mathbb{I}$ which is unity on the truncated prior support $\Gamma_{\theta_i}^{R}$ and zero otherwise, and the normalizing constant Z which can be interpreted as the fractional volume of the truncated prior.

One of the drawbacks of using this box truncation scheme is that it neglects parameter correlations. For higher ($d \gtrsim 2$) dimensional marginal posteriors, this results in the new constrained region usually containing significantly more probability mass than is actually required (see e.g. Ref. [81]).

In this section, we propose a parameter block-wise truncation scheme ("correlated" truncation) based on high-dimensional ratio estimators that accounts for correlations between parameters. This correlated truncation region is defined through a hard likelihood-to-evidence ratio constraint,

$$\tilde{p}_{\Gamma}^{(R)}(\boldsymbol{\theta}) = \frac{1}{Z}\mathbb{I}(\boldsymbol{\theta} \in \Gamma_{\boldsymbol{\theta}}^{(R-1)}) p(\boldsymbol{\theta})\,, \tag{4.7}$$

where $\Gamma_{\boldsymbol{\theta}}^{(R-1)}$ is as defined in Equation (2.27). In Figure 4.1, we show a visual comparison of the two prior truncation strategies in a simple 2-dimensional parameter space. However, following this approach brings about a new challenge that will be addressed in the next section: how does one efficiently define the boundaries of $\Gamma_{\boldsymbol{\theta}}^{(R)}$ and sample from within it?

4.2.3 Sampling from ratio estimators: posterior and constrained prior samples

As discussed in Section 4.1, NRE does not have sampling functionalities and Monte Carlo-based sampling algorithms are used to obtain samples from the approximate posterior.[5] Moreover, we are not only interested in posterior samples, but also in how we can efficiently draw constrained prior samples from a region defined through a hard likelihood-to-evidence ratio constraint. Here, we propose utilizing a *slice-based nested sampling algorithm*, as an efficient method to sample both posterior samples and

[5] It is worth noting that this is somewhat contrary to NPE, where one can directly draw posterior samples from the normalising flow.

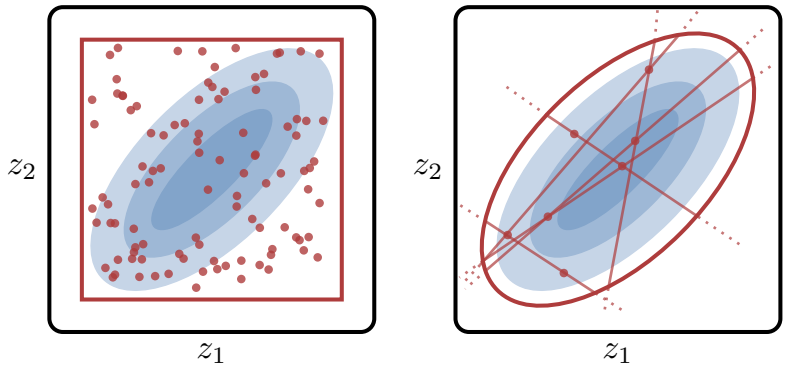

Figure 4.1: Prior truncation strategies. We visualize in a 2-dimensional parameter space a constrained prior region using hyper-rectangular boxes as defined in Equation (4.6) (red rectangle on the left), and a constrained prior region using hard iso-likelihood-to-evidence ratio contours as defined in Equation (4.7) (red ellipse on the right). Sampling from hyper-rectangular boxes is equivalent to uniform sampling, whereas sampling from iso-likelihood-to-evidence ratio contours requires more sophisticated sampling algorithms. In particular, we use a slice-based nested sampler as explained in Section 4.2.3.

constrained prior samples from the ratio estimator. Importantly, constrained prior samples can be obtained through the same nested sampling techniques. Our sampler choice is strongly inspired by the success of slice-based nested samplers in efficiently and reliably exploring high dimensional parameter spaces [100, 294]. Additionally, slice sampling was proposed in Ref. [149] as a sampler for NLE.

First proposed in Ref. [95], nested sampling allows one to sample from high-dimensional probability densities by evolving an ensemble of live points through a high-dimensional parameter space. At the core of nested sampling algorithms is the problem of constrained prior sampling within iso-likelihood contours [97]. This opens up the possibility to re-use technology developed for nested sampling for the purpose of sampling not only from the posterior, but also from the constrained prior region, as defined in Equation (2.27) within iso-likelihood-to-evidence ratio contours.[6]

Typically, nested sampling algorithms start by drawing a collection of live points from the prior. In the current context, one evolves them by discarding the point with the lowest ratio, denoted with $r_{\min}$, and replacing it with a new point subject to the constrain $r > r_{\min}$. The remaining live points are now uniformly distributed over a compressed volume (as they are drawn from a constrained prior as in Equation (2.27)

[6] In traditional implementations the contours are defined by iso-likelihood levels, not iso-likelihood-to-evidence ratio levels.

with $\epsilon = r_{\min}$).

The most challenging aspect of the nested sampling algorithm is drawing new live points under the hard likelihood-to-evidence ratio condition $r > r_{\min}$. One possible reliable and efficient way to do so is through slice sampling, first introduced in Ref. [99]. Starting from one randomly chosen live point, slice sampling builds a chain of proposed live points by taking sequential 1-dimensional steps in a random direction. The length of this chain controls the amount of correlation between the new live point and the initial one (for further information see Ref. [295]).

In our slice sampling implementation, we use a number of different ploys to make our sampling scheme more efficient for the purposes at hand. Firstly, instead of a single chain, the user can define the number of slice sampling chains to draw new live points that will be run in parallel. Secondly, the sampler is implemented such that it allows vectorized evaluations of the natively GPU-based ratio estimator, which provides considerable computational speed-up. We also provide useful functionality to compute the threshold ϵ that defines the constrained prior region in Equation (2.27), based on how much probability mass from the current approximate likelihood-to-evidence ratio one wants the region to include. This is relevant for setting the convergence criterion related to posterior mass that defines the iso-likelihood-to-ratio contour, as in Equation (2.27). Importantly, inference errors caused by truncation arise when not enough posterior mass is included in the truncated regions, resulting in wrongly excluded parts of the parameter space of interest (for a detailed discussion and test of the impact that overly tight complex truncation induces on inference, we refer the reader to Appendix 6.10 of Ref. [168]). For sequential SBI applications, we advise the user to conservatively choose the threshold ϵ such that the truncated prior encloses the highest probability density (HPD) region that contains at minimum 99.9% (i.e. $1 - 10^{-3}$) of the probability mass.

Network stability during sampling. During the sampling step, numerical issues can arise because the slice sampler may evaluate the network outside of the current truncation region $\Gamma_{\boldsymbol{\theta}}^{(R)}$ when searching for new points. In order to detail the issue, we recall our definition of $B = p(\boldsymbol{\theta})/\prod_{i=1}^{d} p(\theta_i)$, which is zero outside of the support of the joined distribution $p(\boldsymbol{\theta})$. Formally, this is not a problem for the ratio A/B (see Equation (4.5)), because that region coincides with the prior region that we truncated away in our truncation procedure (i.e. regions of very low posterior mass). In practice, during sampling, since the network in a given round R has not seen training samples outside of the constrained prior $\tilde{p}^{(R)}(\boldsymbol{\theta})$ (see Equation (4.7)), this can lead to estimated values of our neural likelihood-to-evidence ratio A/B which are much larger than the formal expectation (indeed $B = 0$ exactly outside this region).

We solve this stability issue with the following strategy, that has no impact within

the truncation bounds $\Gamma_{\boldsymbol{\theta}}^{(R)}$, and therefore on the results. During the sampling step, the following substitution is performed in the network

$$B \to \max(B, \min_{\Gamma_{\boldsymbol{\theta}}^{(R)}} B). \tag{4.8}$$

Hence, B can be at minimum as small as the smallest value it can take inside the truncation bound $\Gamma_{\boldsymbol{\theta}}^{(R)}$. As a result, the substitution will never happen inside the constrained region $\Gamma_{\boldsymbol{\theta}}^{(R)}$ where the network was trained, and consequently will not have any impact on the new truncation set $\Gamma_{\boldsymbol{\theta}}^{(R+1)} \in \Gamma_{\boldsymbol{\theta}}^{(R)}$. We further emphasize that by construction, the truncation region should never expand throughout a round, so this new substitution only plays the role of fixing the stability issues outside the constrained prior.

To summarise, in this section we have presented the main contributions of this work in the form of two new building blocks of the TMNRE framework: autoregressive NRE and a slice-based nested sampler to sample posterior and constrained prior samples from a ratio estimator. In the next section, we will apply these newly developed tools to a toy Gaussian example, as well as two astrophysical applications – stellar streams and strong lensing.

4.3 Experiments

We first apply our newly developed tools to a toy multivariate Gaussian example. For more realistic settings, we consider two astrophysics problems: stellar streams parameter inference and substructure parameter inference in strong gravitational lensing observations.

4.3.1 Multivariate Gaussian: scalability

In this first application, we are interested in testing how the performance of "vanilla" NRE and autoregressive NRE models scale with dimensionality and simulation budget. To do so, we consider a d-dimensional Gaussian toy model with strong correlations. The parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ are drawn from a multivariate normal prior $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_0})$, with diagonal covariance matrix $\boldsymbol{\Sigma_0} = 0.1 \odot \mathbf{I}$. The observations $\boldsymbol{x} \in \mathbb{R}^d$ are drawn from a multivariate normal likelihood $p(\boldsymbol{x} \mid \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, with fixed covariance $\boldsymbol{\Sigma}$, which has correlation scales of 0.1 for the off-diagonal entries, and 0.11 for the diagonal ones. As a reference, we use the analytic solution for the true posterior given by Bayes' theorem.

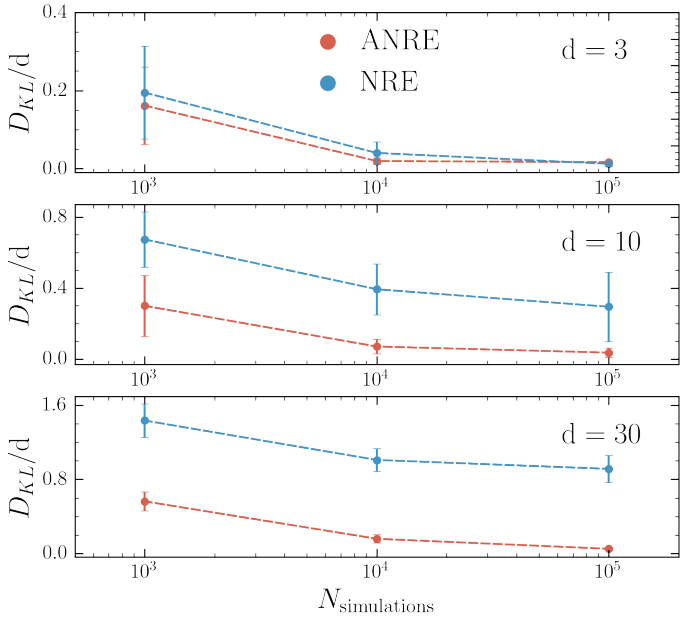

Figure 4.2: Multivariate Gaussian toy example results. In the three panels we show the KL divergence value between the analytical posterior distribution and the estimated ones from the autoregressive (ANRE, in red) a non-autoregressive (NRE, in blue) models. The metric is shown as a function of simulation budget in three different dimensionalities ($d = 3, 10, 30$). For our fixed network capacity, NRE is able to perform as well as ANRE for high simulation budget in low-dimensional scenarios. For higher-dimensionality, the quality of posterior samples obtained from the autoregressive model matches significantly more closely the one from the analytic solution based on the KL divergence value.

We test NRE and autoregressive NRE for dimensions $d = 3, 10, 30$ and for different simulation budgets of $N_{\text{simulations}} = 10^3, 10^4, 10^5$ training data examples. Each conditional ratio in the autoregressive network is modeled with 4 ResNet [296] blocks with 64 hidden features (implemented in `swyft`). The "vanilla" NRE network is instead modeled with a single network whose number of ResNet blocks and parameters is adjusted accordingly to match the total number of parameters in the autoregressive model.

To compare the performance of the two models we consider the Kullback-Leibler (KL) [155] divergence D_{KL} between the posterior estimated by each model and the analytical solution. The KL divergence is a statistical distance that measures the difference between two probability distributions; the closer it is to zero, the better agreement there is. The mean and the covariance of the estimated posterior are computed from the posterior samples of the two models, obtained with our slice-based nested sampler. Our results are shown in Figure 4.2, where each point is the mean KL divergence of 10 different observations. The error bar represents the standard deviation of these values. In low-dimensional scenarios ($d = 3$), NRE is able to perform as well as autoregressive NRE for high simulation budget. For higher-dimensionality, the quality of posterior samples obtained from the autoregressive model matches significantly more closely the one from the analytic solution for lower simulation budgets. In Appendix 4.B, we show the comparison between the analytical posterior and the estimated posterior with simulation budget $N_{\text{simulations}} = 10^5$ from the two models for one observation with $d = 10$.

4.3.2 Stellar streams: autoregressive variable ordering

Stellar streams are very old, dynamical objects in the Milky Way that form as the result of the tidal disruption of objects such as globular clusters or small dwarf galaxies as they orbit the host. In principle, they can act as detailed tracers of the galactic potential, the disruption history and physics of the star clusters, and even possible dark substructures in the Milky Way (see e.g. Refs. [297–304] for examples of the existing analyses). On the other hand, robust analysis of these objects is challenging for a number of reasons. Firstly, simulating the stream in full generality is computationally challenging, so delicate modeling choices need to be made. Secondly, the statistical properties of the resulting stream are difficult to write down explicitly in the form of a likelihood, mainly due to complicated observational effects such as membership probabilities in e.g. Gaia data [305, 306]. Combined with the general simulation efficiency that has been observed with neural ratio estimation, this strongly motivates the application of SBI techniques to such a problem, something that has been studied in detail in Ref. [90] (see also Ref. [89]).

In the current context, we will use the inference of stream model parameters in

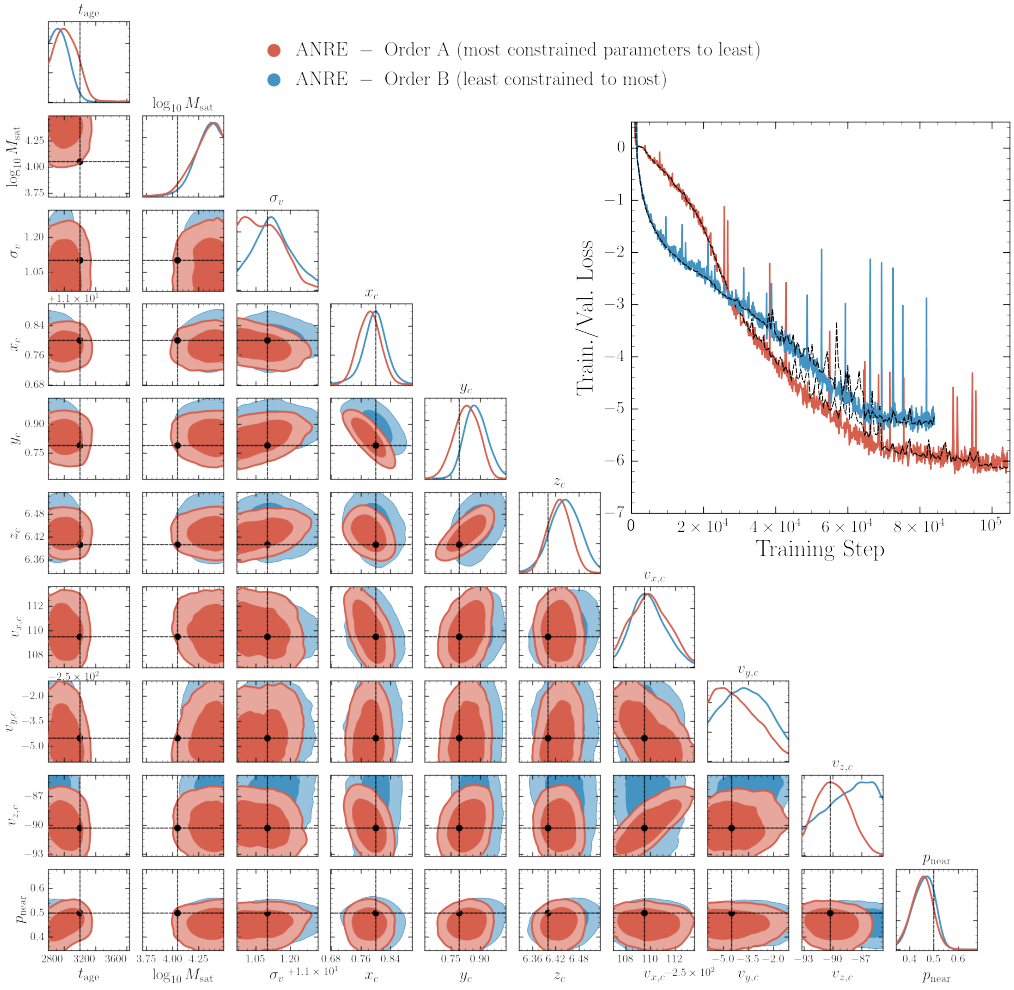

Figure 4.3: Stellar streams experiment results. *Corner plot:* The corner plot shows the comparison between the inference results using different parameter ordering schemes across those parameters that are constrained in the analysis (for the full set, see Appendix 4.C). Order A (most constrained parameters to least constrained) is shown with red contours, whilst Order B (least constrained to most) is shown in blue. We see that Order A performs marginally better in terms of precision, although the improvement is somewhat marginal. The true injected values are shown with black dashed lines and dots. *Upper right inset:* This inset shows the training loss (see Equation (2.21)) as a function of the number of training steps for the two orderings. We see that Order A achieves a lower value of the loss, in line with the conclusions from the corner plot. The validation loss for each case is shown with black dashed lines.

an identical set up to Ref. [90] to highlight a relevant aspect of the autoregressive algorithm presented in this work. Specifically, we use the case study to highlight the implications that varying the ordering of parameters in the conditional distributions product found in Equation (4.3) has on inference performance. The intuition that we wish to test is the following: we expect that the distribution $p(\theta_i|\boldsymbol{x})$ is most useful to learn if θ_i is well constrained. Similarly, if θ_j is not well constrained, then adding conditional information e.g. in $p(\theta_j|\boldsymbol{x}, \boldsymbol{\theta}_{1:j-1})$ can add relevant information. With this said, the expectation is that an approximately optimal ordering runs from θ_1 as the most constrained parameter (relative to the prior), and θ_d as the least constrained parameter.

Given the analysis in Ref. [90], we have good expectations about the relative constraining power of the streams analysis on various parameters. As such, we choose two orderings: Order A which runs from most to least constrained, and Order B which runs vice versa (see Appendix 4.C for details). We train an autoregressive estimator in both cases to perform inference on all parameters of the model. The results of this experiment are shown in Figure 4.3 in terms of the resulting posterior distributions, as well as the training and validation loss curves for the neural network training process.

There are two things to note from Figure 4.3: firstly, the ordering of parameters does have an effect on the training procedure, with the expectation laid out above clearly realised. In other words, we see a plateauing of the optimal training and validation loss for Order B at a value that is clearly larger than the asymptotic value in the case of Order A. In addition, broadly the overall quality of posterior inference on parts of the model that are well measured is of higher quality with Order A. Together, this suggests that the general intuition regarding parameter ordering in autoregressive models holds here also. With that being said, however, we do observe that the 'penalty' for choosing a non-optimal ordering is not severe, at least in this case, and high fidelity inference could be achieved by e.g. increasing the simulation budget (see Appendix 4.C for an example), or following the prior truncation scheme described above and iterating the simulation generation, training, and inference steps. To be clear on this last point, this example is designed to demonstrate for fixed simulation budget and fixed network architecture the impact of variable ordering. We expect in this particular case that if the prior truncation scheme is followed, the posteriors will continue to shrink and converge to the same answer. For high precision inference, we advise that on convergence, an explicit test varying the ordering is carried out to confirm that the posteriors do not shift by more than the desired sensitivity. For full technical details regarding this example, see Ref. [90] or Appendix 4.C. Also in Appendix 4.C, we show two other random orderings to highlight the relative insensitivity to this choice and build on the example in subsection 4.3.1.

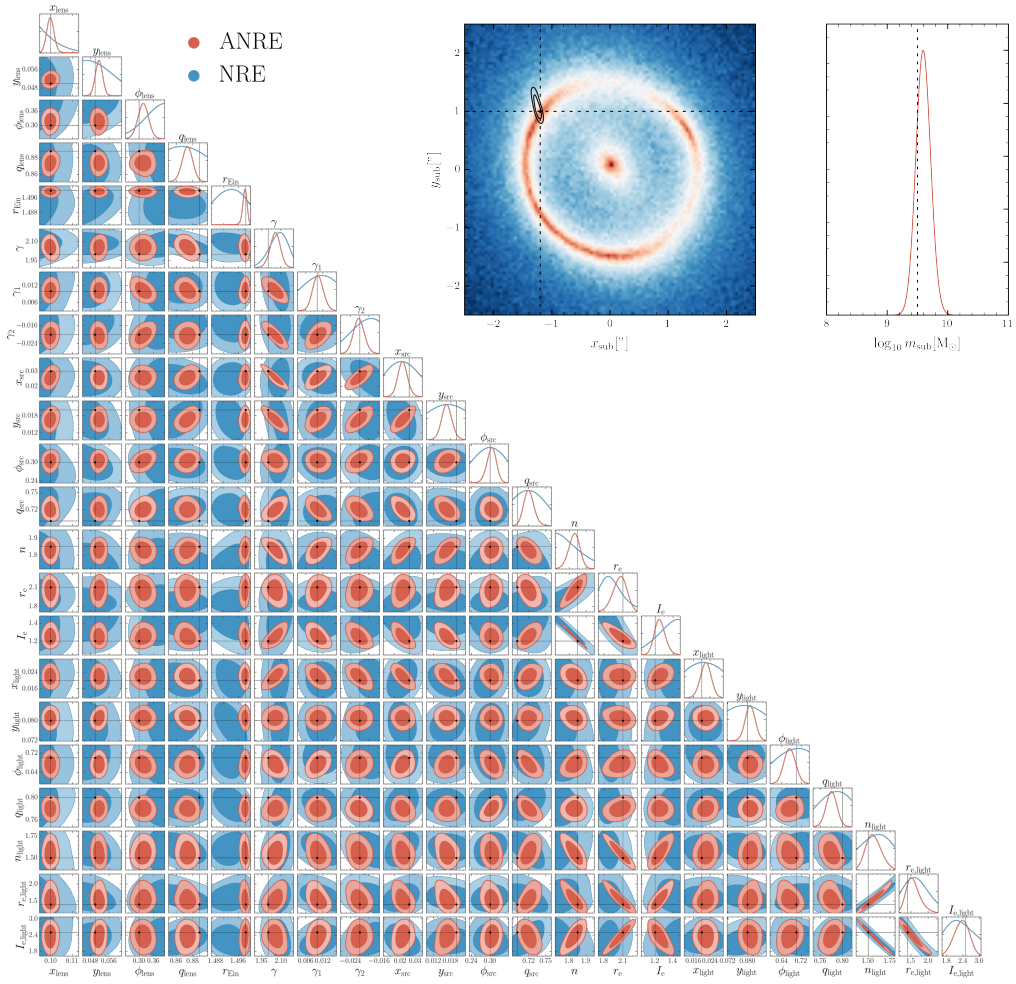

Figure 4.4: Strong gravitational lensing experiment results. *Corner plot:* The corner plot shows the comparison between the macro-model parameter inference results using the autoregressive NRE model (ANRE, red contours) and the non-autoregressive one (NRE, blue contours). The true values are shown with black dashed lines and dots and are shown in Table 4.D.1. *Upper right inset:* The inset shows the results for subhalo parameter inference for our target mock observation.

4.3.3 Strong gravitational lensing: correlated truncation

As described in Chapter 3, galaxy-galaxy strong gravitational lensing occurs when the paths of light rays from a background galaxy are distorted by the mass of an intervening lens galaxy before entering a telescope [27]. This leads to extremely distorted ring-shaped images with multiple copies of the source. If a small-scale dark matter halo is present in such a system, its distortions will be localized, mostly impacting the image of the source along the nearest line of sight. As first proposed in Ref [201], by carefully analyzing the relationship between the multiple images of the source, the distortions due to a dark matter perturber can be disentangled from possible variations in the source light and its properties can be measured.

In reality, substructure parameter inference in strong gravitational lensing systems is a very challenging task for a number of reasons. In particular, the desired signal manifests as percent level variations, lensing systems are remarkably complicated, and the variance between images is high (i.e. observations are very diverse). A targeted SBI approach to this "needle in a haystack" type of analysis is therefore well-motivated and has been proven successful, see e.g. Refs. [5, 244].

For this application, we consider a lensing system with the following components: source light, lens light, lens mass distribution, and external shear (which we collectively call the macro-model); and one single subhalo. More details on the system and noise model can be found in Appendix 4.D. The blending effect due to the overlap of the light emitted by the lensed source and by the lens itself complicates the interpretation and analysis of observed lensed systems, and makes it challenging to isolate and study the properties of the lensed source. This blending effect can be mitigated if multi-wavelength observations are available. Usually the lens light gets subtracted assuming the best-fit value (see e.g. Ref. [37]). In this SBI application, we leave it free to vary and infer its parameters at the same time as all other components, accounting in the analysis for its uncertainties and the correlations it has with the rest of the system components.

We perform inference on the simulated target observation shown in Figure 4.4, generated with the parameters in Table 4.D.1. The first step in the analysis is to reduce training data variance by constraining the macro-model parameters prior. Starting from the full macro-model prior shown in Table 4.D.1, we use the procedure described in Refs. [1, 5] and Chapter 3 to constrain it sequentially via TMNRE rounds using box truncation. When convergence[7] is reached, we are not able to reduce any further the data variance displayed by the simulations via 1-dimensional marginal posterior estimation and box truncation.

[7] We define the ratio estimator to be converged when, after two consecutive rounds in which we double up the simulation budget, the constrained hyper-rectangular box prior has shrunk by less than 10%.

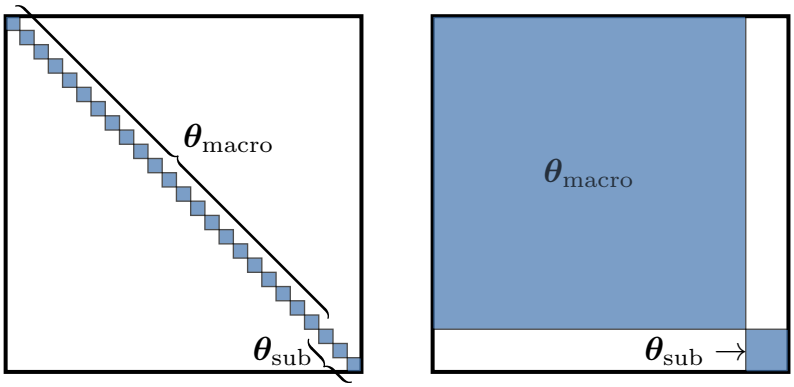

Figure 4.5: Prior truncation strategies. *Left:* With a parameter-wise truncation strategy, each parameter gets individually constrained with boxes. *Right:* Using a parameter block-wise truncation strategy, we divide the parameter space in blocks (lensing macro-model parameters $\boldsymbol{\theta}_{\text{macro}}$ and subhalo parameters $\boldsymbol{\theta}_{\text{sub}}$), depending on what dominates the data variance, and which parameters are most correlated. During sequential inference, we account for correlations between parameters in the same block through correlated truncation.

As previously discussed in Section 4.2.2, a disadvantage of using hyper-rectangular boxes lies in the fact that for high-dimensional and correlated parameter spaces the constrained region contains more probability mass than required. We will now employ a parameter block-wise correlated prior truncation strategy instead of box truncation, as presented in Section 4.2.2. In Figure 4.5, we show a simple visualization of the two prior truncation strategies for this specific application.

In order to learn the macro-model parameter correlations, especially the ones between lens light and source light, we model the joint posterior of the macro-model parameters. We train two models to estimate the macro-model joint ratio: a "vanilla" NRE model and the autoregressive NRE model presented in Section 4.2.1. We used the same simulation budget and amount of network weights (for more details regarding training and the employed neural network architecture, see Appendix 4.D). In Figure 4.4, we compare the posterior samples obtained from the two models using our slice-based nested sampler presented in Section 4.2.3. It is clear that, for this application and simulation budget, autoregressive NRE performs significantly better than NRE, which is not able to properly model the joint macro-model parameter distribution.

We then sample correlated constrained prior samples with our slice-based nested sampler. By modeling and exploiting the intricate correlations between lens light, source light, and lens parameters, instead of using a parameter-wise truncation strat-

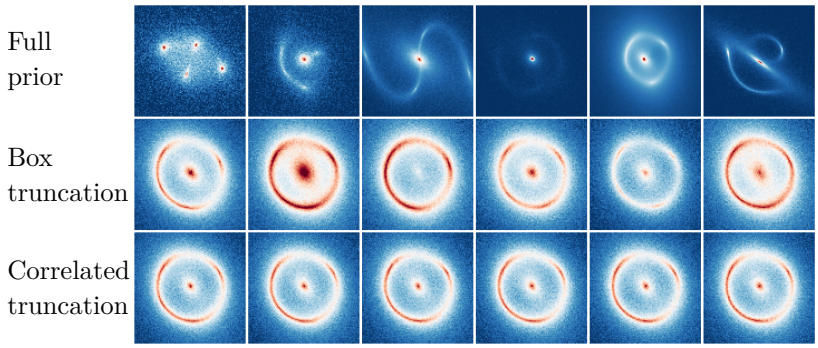

Figure 4.6: Targeted training data. The simulations in the first row are drawn from the full initial prior, described in Table 4.D.1. In the intermediate row, we show simulations used for the last round of training of the 1-dimensional marginal ratio estimators, where the macro-model parameters were drawn from the most constrained hyper-rectangular box prior. In the last row, the macro-model parameters were sampled from the high-dimensional constrained prior with the slice-based nested sampler. The images in the second and last row are plotted in the same color scale, whereas the ones in the first row are allowed to vary for visualization purposes since they exhibit a larger variance.

egy, we are able to achieve a much lower data variance. In Figure 4.6, we display examples of training data. The simulations in the first row are drawn from the full initial prior. In the intermediate row, we used macro-model parameters drawn using the box truncation scheme. Finally, the simulations presented in the last row use the slice-based nested sampler to sample the macro-model parameters from the high-dimensional constrained prior, using the autoregressive model.

Having reduced the macro-model parameters training data variance, we are then able to focus on substructure parameter inference, for which the results are shown in Figure 4.4. These results were obtained with two sequential rounds of inference using a correlated truncation scheme, as first explained in Section 2.2.3. In contrast, given the initial fixed simulation budget, we were not able to constrain the subhalo parameter in our target observation with the data variance displayed by simulations using only box truncation.

4.4 Discussion and conclusions

The motivation for the SBI method development presented in this work comes from the data analysis and statistics challenges facing the fields of astrophysics and cosmol-

ogy in light of high quality data from current and future experiments [10, 12, 42, 276–279]. In this work, we have focused on one specific SBI algorithm, TMNRE, and advanced its potential with the introduction of two new components:

a) *Autoregressive NRE*, presented in Section 4.2.1, estimates multi-dimensional (marginal) posteriors by effectively computing d 1-dimensional conditional ratios instead of a d-dimensional joint one (see Equation (4.1)). This way of approaching high-dimensional distribution estimation through an autoregressive scheme has been motivated by its proven scaling potential in other SBI techniques [287–289]. We release a publicly available implementation of the autoregressive model in the TMNRE package `swyft`[8].

b) Our *slice-based nested sampling* implementation, presented in Section 4.2.3, offers vectorized evaluations of the natively GPU-based ratio estimator and introduces a convergence criterion related to posterior mass. The latter is essential for sequential inference applications since it allows one to define the truncation region in Equation (2.27). The sampler choice naturally comes from the realisation that nested sampling techniques provide a tool to efficiently sample not only from a multi-dimensional posterior, but also natively from a constrained prior. Our slice sampler implementation is available in the package `torchns`[9].

We have demonstrated their application in three case studies presented in Section 4.3. Firstly, we tested the performance of the autoregressive ratio model against the non-autoregressive one in a toy example. The main results, for different simulation budgets and dimensionality, are presented in Figure 4.2. Secondly, we explored how variable ordering impacts the autoregressive model in a stellar streams analysis. We have found that a non-optimal variable ordering does have an impact, but does not severely penalise the overall inference result in our specific application, as shown in Figure 4.3. Lastly, we investigated the potential of a correlated truncation scheme in a proof-of-concept application to substructure searches in a strong gravitational lensing image. The main outcomes are exhibited in Figs. 4.4 and 4.6: in particular, given our fixed simulation budget, the non-autoregressive model is not able to estimate the macro-model parameter joint distribution, and only through correlated prior truncation are we able to constrain the subhalo parameters.

Outlook. There are two aspects to consider as far as outlook is concerned for our method development. The first concerns the known limitations of the autoregressive modeling and the nested sampling techniques. The second is more general, and considers the analysis settings in which this approach may be useful.

[8] https://github.com/undark-lab/swyft
[9] https://github.com/undark-lab/torchns

As discussed above, one of the potential limitations of autoregressive modeling is its sensitivity to parameter orderings. Although we investigated this in the context of stellar streams in Section 4.3.2, and found that the effect was minimal, it is possible that this is model dependent. As such, to extend the method, it would be interesting to develop techniques to derive an (approximately) optimal ordering of the parameters, e.g. by estimating the strength of correlations between various parameter sets.

On the nested sampling side, known limitations include the fact that it becomes very costly in terms of the required number of network evaluations for high dimensional parameter spaces, say $d \gtrsim 30$ (for a discussion see, e.g., Ref. [295]). As such, it will be interesting to further investigate methods that exploit gradient information and scale better to high dimensions, such as the proximal nested sampling technique proposed in Ref. [307]. Moreover, our current implementation of the slice sampler does not include clustering detectors (see e.g. Ref. [308]), which makes it inefficient for multi-modal posteriors.

More generally, we believe the method we have developed could be useful in the following type of scenario: suppose that a part of the parameter space dominates the data variance, but is not the one that we are ultimately most interested in performing inference on. Using traditional methods, there are typically two approaches: a) solve the full joint analysis problem for all parameters, but face significant computational challenges, or b) analyse the nuisance parameters separately, and then fix them to some form of best-fit value to use in the rest of the analysis, at the cost of neglecting their uncertainties. For example, as briefly mentioned in Section 4.3.3, this is exactly the scenario in standard strong gravitational lensing analyses where the lens light gets subtracted assuming the best-fit value, before analysing the lensed emission.

With the method developed in this work, however, there is an alternative approach available. In particular, as explained in Section 2.2.3, the fact that prior truncation composes well with marginalisation allows us to consistently combine different analysis strategies that target distinct parts, or "blocks", of the model into a coherent "inference assembly". This exact procedure was exemplified in Section 4.3.3, where we first reduced training data variance by truncating the lens and lensed source parameters, and then performed sequential inference on our parameters of interest, the ones pertaining the substructure.

In conclusion, this approach can be relevant for many different cosmological and astrophysical applications, where the analysis would be carried out simultaneously on different components of the systems, e.g. parameters of interest, foregrounds, backgrounds, nuisance parameters, instrumentation parameters *etc.* Ultimately, this will allow us to flexibly combine different inference strategies in order to draw coherent and consistent conclusions based on the full model and all the data.

Appendix

4.A Another formulation

The autoregressive model presented in Section 4.2.1 can be also written with a single network that can estimate both components. In order to do this, we introduce an auxiliary variable $c = -1, 1$. We then consider the ratio

$$r(\theta_i; \boldsymbol{x}, c, \boldsymbol{\theta}_{1:i-1}) = \frac{p'(\theta_i \mid \boldsymbol{x}, c, \boldsymbol{\theta}_{1:i-1})}{p(\theta_i)} \tag{4.9}$$

The distribution of our auxiliary model is now given by

$$\theta_i, \boldsymbol{x}, c, \boldsymbol{\theta}_{1:i-1} \sim p'(\boldsymbol{x}, c, \theta_i, \boldsymbol{\theta}_{1:i-1}) \equiv p'(\boldsymbol{x} \mid c, \theta_i, \boldsymbol{\theta}_{1:i-1}) p(c) p(\boldsymbol{\theta}_{1:i}) \ , \tag{4.10}$$

where we make use of the definitions $p'(\boldsymbol{x} \mid c = -1, \theta_i, \boldsymbol{\theta}_{1:i-1}) \equiv p(\boldsymbol{x})$ and $p'(\boldsymbol{x} \mid c = 1, \theta_i, \boldsymbol{\theta}_{1:i-1}) \equiv p(\boldsymbol{x} \mid \theta_i, \boldsymbol{\theta}_{1:i-1})$.

Once the above ratio estimator is trained, we can obtain the desired conditional posterior-to-conditional prior ratio by using

$$\frac{p(\theta_i \mid \boldsymbol{x}, \boldsymbol{\theta}_{1:i-1})}{p(\theta_i \mid \boldsymbol{\theta}_{1:i-1})} \approx \frac{r(\theta_i; \boldsymbol{x}, c = +1, \boldsymbol{\theta}_{1:i-1})}{r(\theta_i; \boldsymbol{x}, c = -1, \boldsymbol{\theta}_{1:i-1})} \ . \tag{4.11}$$

The advantage of this formulation is that it reduces the number of networks to train by a factor of two, which should correspondingly reduce GPU memory requirements and training time. A potential downside of this approach is that the network capacity has to be high enough in order to efficiently learn both posterior and prior approximations at the same time. We will leave a quantitative comparison between different methods and network architectures to future work.

4.B Multivariate Gaussian experiment

In Figure 4.B.1, we show the agreement between the analytical and estimated posteriors for the NRE and ANRE methods applied to the toy problem described in subsection 4.3.1. We see that across all 10 parameters the autoregressive model achieves almost perfect agreement with the analytic result.

In order to check the impact of the autoregressive ordering in this Multivariate Gaussian toy example we have conducted an additional test for 10 dimensions. For this test, we use a likelihood with fixed covariance $\boldsymbol{\Sigma}$, which has correlation scales of 0.1 for the off-diagonal entries, and varying diagonal correlation scales, ranging from

Parameter	True value	Initial prior
t_{age} [Myr]	3000	$\mathcal{U}(2583, 3647)$
λ_{rel}	1.405	$\mathcal{U}(0.13, 2.00)$
λ_{match}	1.846	$\mathcal{U}(0.31, 2.00)$
ξ_0	0.001	$\mathcal{U}(0.0001, 0.01)$
α	20.9	$\mathcal{U}(10.0, 30.0)$
r_h [pc]	0.001	$\mathcal{U}(0.0001, 0.01)$
$\bar{m}$ [$M_\odot$]	3	$\mathcal{U}(1.0, 20.0)$
$\log(M_{\text{sat}}/M_\odot)$	4.05	$\mathcal{U}(3.3, 4.5)$
σ_v [km/s]	1.1	$\mathcal{U}(0.96, 1.32)$
x_c [kpc]	11.8	$\mathcal{U}(117, 119)$
y_c [kpc]	0.79	$\mathcal{U}(0.6, 1.1)$
z_c [kpc]	6.4	$\mathcal{U}(6.32, 6.54)$
$v_{x,c}$ [km/s]	109.5	$\mathcal{U}(106.8, 113.7)$
$v_{y,c}$ [km/s]	-254.5	$\mathcal{U}(-256, -251)$
$v_{z,c}$ [km/s]	-90.3	$\mathcal{U}(-93.3, -84.6)$
p_{near}	0.5	$\mathcal{U}(0.34, 0.72)$

Table 4.B.1: True parameter values and priors used in the TMNRE inference round for the stellar streams example. Note that our prior choices are taken from the final round of inference of Ref. [90] so as to test the autoregressive model in a final round of precision inference through active learning.

0.1 to 0.55 in 0.05 steps. We have run our autoregressive model for four different variable orderings: from most constrained to least, from least constrained to most, and two random orderings. The results shown in Figure 4.B.2 demonstrate that the autoregressive ordering does not have an impact in this case.

4.C Stellar streams experiment

In this short appendix we present the technical details relevant for the stellar streams experiment presented in Section 4.3.2. There are a number of things directly analogous to the implementation described in Ref. [90]. In particular, we use the stellar streams simulator, `sstrax` [108], developed in this reference, as well as the `swyft`-based inference code `albatross` [309]. In addition, we consider the same set of model parameters (described in Table 4.B.1 and in detail in Ref. [90]), the same injection

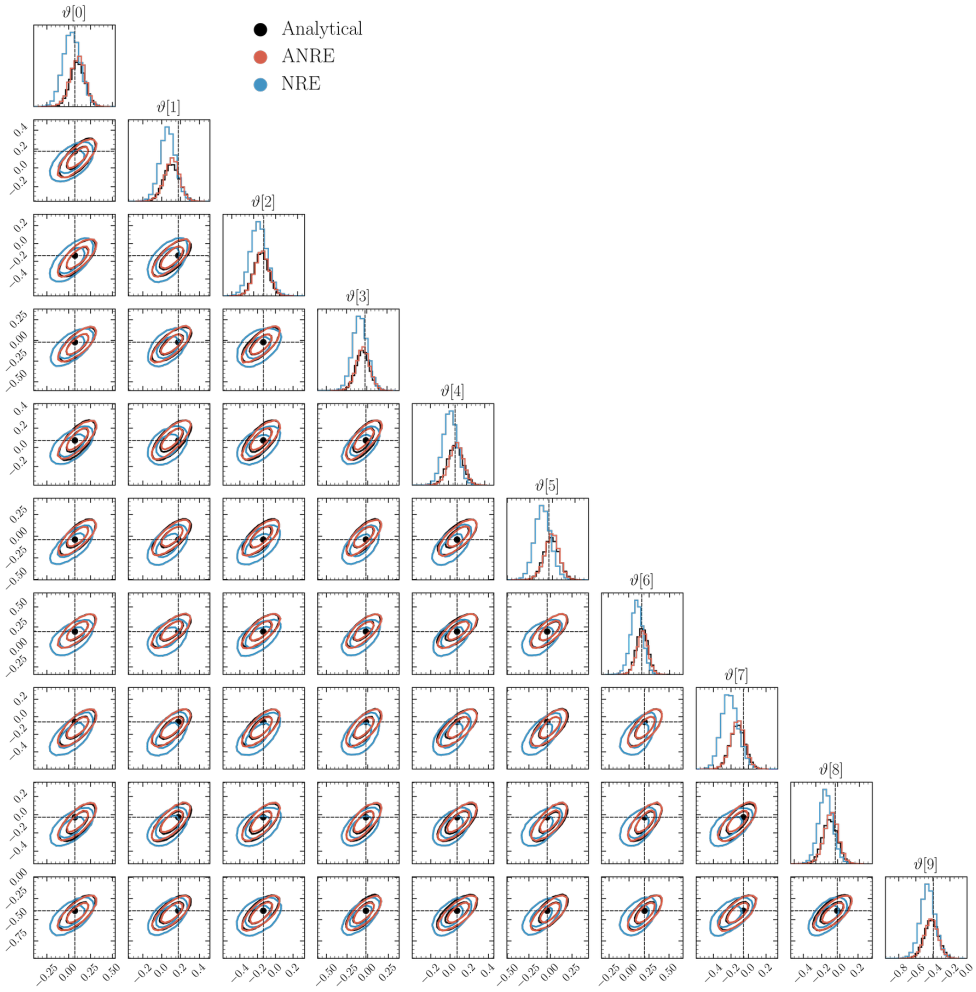

Figure 4.B.1: Multivariate Gaussian toy example results. Corner plot highlighting the agreement between the analytic and estimated posteriors for the $d = 10$ Gaussian model described in subsection 4.3.1.

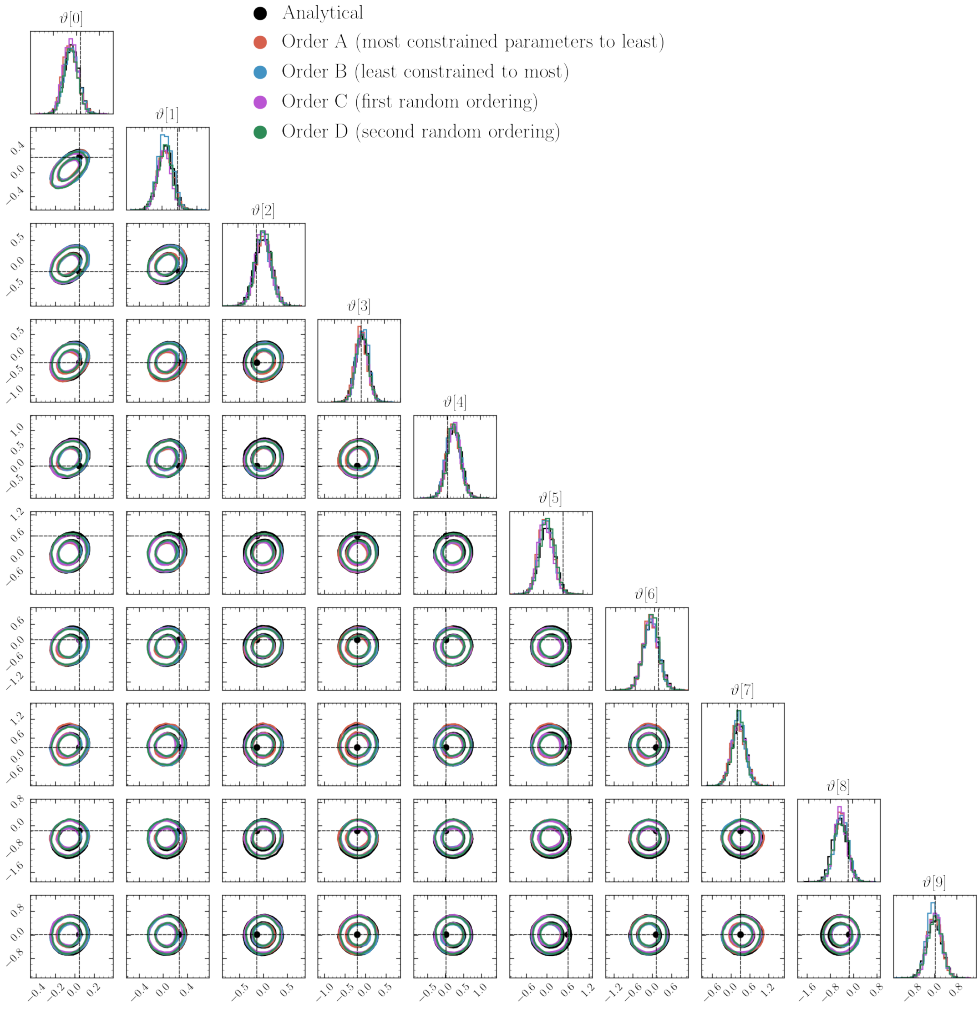

Figure 4.B.2: Multivariate Gaussian toy example results. Corner plot highlighting the agreement between the analytic and estimated posteriors for the $d = 10$ Gaussian model described in subsection 4.3.1 and Appendix 4.B. For the estimated posteriors we have used four different orderings, as described in Appendix 4.B.

parameters to generate the observation, and the same binning/observation scheme. The key difference with respect to this reference is in the inference network used. In particular, we replace the 1-dimensional ratio estimation across all model parameters with the autoregressive method described in this work to model their joint distribution. In terms of network training details, we use the Adam optimiser with an initial learning rate of 10^{-6}. In addition, we use a batch size of 512 and a simulation budget of 3×10^5 training examples. Finally, we use the following orderings (Order A and Order B referenced in the main text) for training the autoregressive ratio estimator: [σ_v, x_c, y_c, z_c, $v_{x,c}$, $v_{y,c}$, $v_{z,c}$, $t_{\rm age}$, $p_{\rm near}$, $\lambda_{\rm rel}$, λ_{match}, ξ_0, α, r_h, $\bar{m}$, $\log(M_{\rm sat}/{\rm M}_\odot)$] (Order A), and [$\xi_0$, λ_{match}, $\lambda_{\rm rel}$, α, r_h, $\bar{m}$, $\log(M_{\rm sat}/{\rm M}_\odot)$, $t_{\rm age}$, σ_v, x_c, y_c, z_c, $v_{x,c}$, $v_{y,c}$, $v_{z,c}$, $p_{\rm near}$] (Order B).

In Figure 4.C.1, we also show the results of a test with two additional random orderings. Specifically, the orderings shown are [α, $\log(M_{\rm sat}/{\rm M}_\odot)$, $\lambda_{\rm rel}$, ξ_0, $v_{x,c}$, σ_v, $v_{z,c}$, $\bar{m}$, $t_{\rm age}$, r_h, x_c, z_c, $p_{\rm near}$, $\lambda_{\rm match}$, y_c, $v_{y,c}$] (first random order) and [$v_{z,c}$, $p_{\rm near}$, $\bar{m}$, σ_v, $v_{y,c}$, α, $v_{x,c}$, $t_{\rm age}$, $\lambda_{\rm rel}$, z_c, $\lambda_{\rm match}$, y_c, x_c, ξ_0, r_h, $\log(M_{\rm sat}/{\rm M}_\odot)$] (second random order).

Furthermore, we performed an additional test in order to check when the loss for order B reaches the one for order A in Figure 4.3 (see main text for definitions), by gradually increasing the simulation budget. Figure 4.C.2 shows the loss for order B as a function of epochs, for different sizes of training dataset, with the biggest one (100% in the figure) containing 6×10^5 simulations. From Figure 4.C.2, we can see that the loss for order B reaches a plateau value of -6.1 (the one of the loss for order A in Figure 4.3 when training on 3×10^5 simulations) for approximately between 60% to 70% of the total 6×10^5 simulations, so roughly for 4×10^5 simulations. In this specific example, this experimental check shows that a network trained with order B needs 25% more simulations to plateau at the same loss as one trained for order A. This test further confirms our observation that the 'penalty' for choosing a non-optimal ordering is not severe, and can be easily circumvented, for example as in this case, by increasing the simulation budget. However, the important take away message is that the simulation computational budget can be reduced by choosing an optimal ordering.

4.D Gravitational lensing experiment

In this appendix we provide a more detailed account of the model adopted to generate the strong lensing simulations, the training details, and the employed neural networks to obtain the results presented in Section 4.3.3.

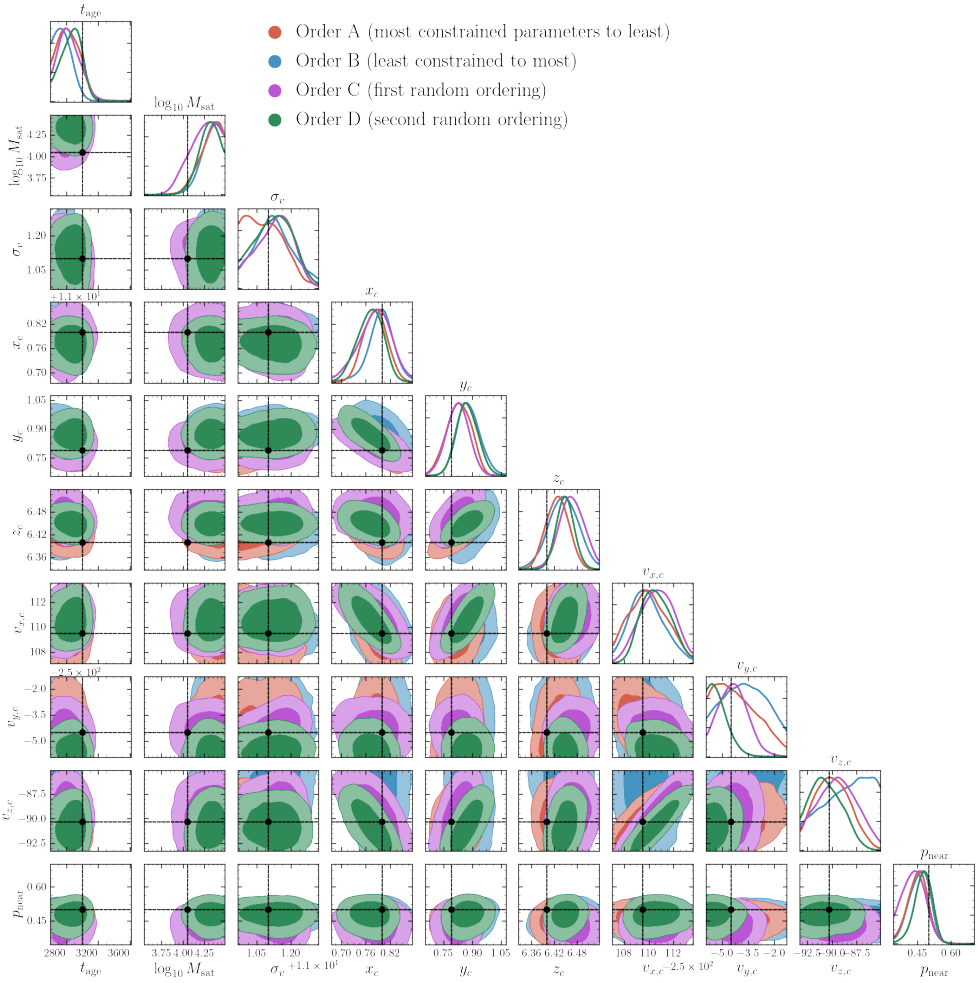

Figure 4.C.1: Additional checks of the variable ordering conclusions given in Section 4.3.2. Here, we overlay the results presented in the main text with inference results given two other random orderings.

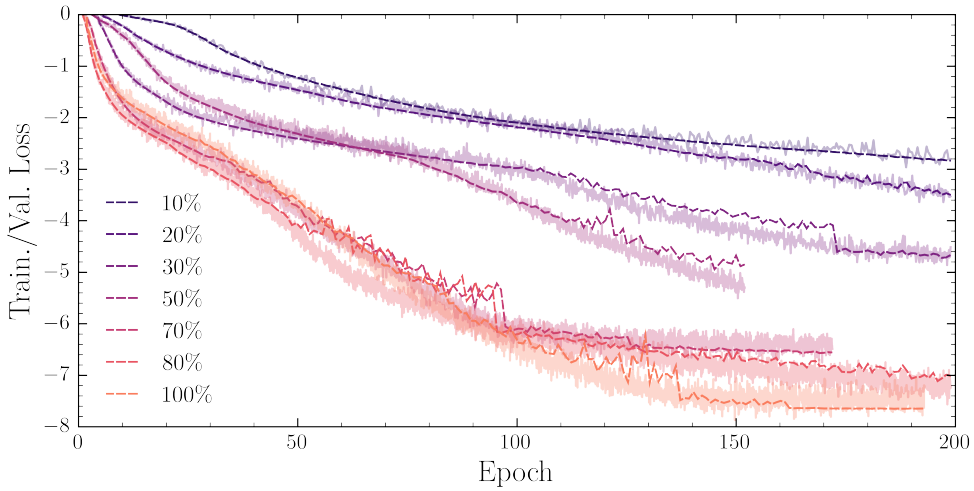

Figure 4.C.2: Training loss for order B (see Figure 4.C.1 and Section 4.3.2 and Figure 4.3 in the main text) as a function of epochs for different simulation budgets, with a maximum budget of 6×10^5 simulations. The validation loss for each case is shown as a dashed line in the figure.

4.D.1 Simulator

We use the simulator adopted in Refs. [1, 5, 244], and described in Section 3.2 to generate strong lensing image observations. To model the lens light and the source light flux we use a Sèrsic profile [251], parameterized by seven variables: position, position angle, axis ratio, index, effective radius, and surface intensity. We adopt a singular power-law ellipsoid (SPLE) [253] for the main lens mass distribution, with a total of six parameters: position on the lens plane, position angle, axis ratio, slope, and Einstein radius. We consider two additional parameters to model the external shear. In total, there are twenty-two macro-model parameters. To model the density profile of the dark matter subhalo we adopt the smoothly truncated universal Navarro-Frenk-White mass density profile from Ref. [261]. We fix the truncation radius to $\tau = 6$. The subhalo is then described by three parameters: virial mass $m_{\rm sub}$, and position on the lens plane $(x_{\rm sub}, y_{\rm sub})$. We adopt the concentration-mass relation from Ref. [310]. We show each parameter prior and the value with which we have generated our mock target observation in Table 4.D.1.

Given this model, we generate 100×100 pixel^2 images with a resolution of 0.05" per pixel side, for a total field of view of $5" \times 5"$ in an image. The instrumental effects include a Gaussian PSF with a full width at half maximum of 0.05" and Gaussian noise. We choose redshift $z_{\rm lens} = 0.9$ for the lens and $z_{\rm source} = 2$ for the source.

4.D.2 Training details and neural networks

For all tasks we use the same general ratio estimator architecture. It consists of an initial compression network that maps the high-dimensional lensing observation $\boldsymbol{x}$ into a feature vector. This feature vector is concatenated to the parameters we want to infer. The vector is then passed to `swyft` binary classifier which outputs an estimate of the likelihood-to-evidence ratio. The compressor architecture for the macro-model parameters is given in Table 4.D.2. This same compression is used when estimating 1-dimensional marginal ratio estimators, or the joint one with NRE or autoregressive NRE, as explained in Section 4.3.3. The compressor architecture for the subhalo parameters is given in Table 4.D.3.

We used the Adam optimizer with an initial learning rate of 10^{-3} for the macro-model ratio estimator and 8×10^{-5} for the subhalo ratio estimator, and a batch size of 64. The learning rate was reduced by a factor of 0.1 whenever the validation loss plateaued for 2 epochs. Training was run until the validation loss stopped improving for more than 5 epochs. The results shown in Figure 4.4 are obtained using 2×10^5 training samples. To run the inference, we use the model weights obtained at the lowest validation loss curve point.

Component	Parameter	True value	Initial prior
Subhalo	$x_{\text{sub}}\,[''] $	-1.2	$\mathcal{U}(-2.5, 2.5)$
	$y_{\text{sub}}\,[''] $	1	$\mathcal{U}(-2.5, 2.5)$
	$\log_{10} m_{\text{sub}}\,[M_\odot]$	9.5	$\mathcal{U}(8, 11)$
SPLE	$x_{\text{lens}}\,[''] $	0.1	$\mathcal{U}(-0.2, 0.2)$
	$y_{\text{lens}}\,[''] $	0.05	$\mathcal{U}(-0.2, 0.2)$
	$\varphi_{\text{lens}}\,[^\circ]$	0.3	$\mathcal{U}(0, 1.5)$
	q_{lens}	0.89	$\mathcal{U}(0.1, 1)$
	γ	2	$\mathcal{U}(1.8, 2.2)$
	$r_{\text{ein}}\,[''] $	1.5	$\mathcal{U}(1, 2)$
Shear	γ_1	0.01	$\mathcal{U}(-0.05, 0.05)$
	γ_2	-0.02	$\mathcal{U}(-0.05, 0.05)$
Lens light	$x_{\text{light}}\,[''] $	0.03	$\mathcal{U}(-0.1, 0.1)$
	$y_{\text{light}}\,[''] $	0.02	$\mathcal{U}(-0.1, 0.1)$
	$\varphi_{\text{light}}\,[^\circ]$	0.3	$\mathcal{U}(0., 1.5)$
	q_{light}	0.7	$\mathcal{U}(0.1, 1)$
	n	1.58	$\mathcal{U}(0.1, 4)$
	$r_e\,[''] $	2.1	$\mathcal{U}(0.1, 3)$
	I_e	1.2	$\mathcal{U}(0, 4)$
Source light	$x_{\text{src}}\,[''] $	0.02	$\mathcal{U}(-0.2, 0.2)$
	$y_{\text{src}}\,[''] $	0.08	$\mathcal{U}(-0.2, 0.2)$
	$\varphi_{\text{src}}\,[^\circ]$	0.7	$\mathcal{U}(0., 1.5)$
	q_{src}	0.8	$\mathcal{U}(0.1, 1)$
	n	1.5	$\mathcal{U}(0.1, 4)$
	$r_e\,[''] $	1.4	$\mathcal{U}(0.1, 3)$
	I_e	2.5	$\mathcal{U}(0, 4)$

Table 4.D.1: True subhalo and macro-model parameter values and priors used in the first TMNRE inference round in the strong gravitational lensing application. This prior generates images with variance displayed in the first row of Figure 4.6.

`Conv2d(1, 4, 3, 2, 1, bias=True)`
`BatchNorm2d(4)`
`ReLU`
`Conv2d(4, 8, 5, 2, 1, bias=True)`
`BatchNorm2d(8)`
`ReLU`
`Conv2d(8, 16, 5, 2, 1, bias=True)`
`BatchNorm2d(16)`
`ReLU`
`Conv2d(16, 32, 5, 2, 1, bias=True)`
`BatchNorm2d(32)`
`ReLU`
`Flatten()`
`LazyLinear(256)`

Table 4.D.2: The convolutional compression network used in the macro-model parameter ratio estimator. The notation is taken from `PyTorch`: the arguments to `Conv2d` are the number of input channels, output channels, kernel size, stride and padding, respectively. The horizontal lines highlight where the number of channels changes. Note that we standardize the images before providing them to the convolutional networks.

`image_size`	100
`n_channels`	1
`n_classes`	8
`s`	1

Table 4.D.3: The details of the UNet network used in the subhalo. We use the implementation from https://github.com/milesial/Pytorch-UNet, with arguments given in the table. The UNet output is then sampled at the positions of the subhalo contrastive examples to bring it to a lower dimensionality, and this feature vector is passed to the binary classifier.

5 Simulation-based inference for cosmological initial conditions

Reconstructing astrophysical and cosmological fields from observations is challenging. It requires accounting for non-linear transformations, mixing of spatial structure, and noise. In contrast, forward simulators that map fields to observations are readily available for many applications. In this chapter, we present a versatile Bayesian field reconstruction algorithm rooted in SBI and enhanced by autoregressive modeling. The proposed technique is applicable to generic (non-differentiable) forward simulators and allows sampling from the posterior for the underlying field. We show first promising results on a proof-of-concept application: the recovery of cosmological initial conditions from late-time density fields.

This chapter is based on work from [3].

5.1 Introduction

Recent developments in simulation-based machine learning are increasingly used for tackling difficult astrophysical and cosmological data analysis challenges [e.g., 1, 2, 4, 5, 55, 60, 64–66, 70, 73, 74, 76, 78, 80, 81, 87, 90, 178, 282, 285, 311–318]. While SBI has primarily been employed to solve relatively low-dimensional ($\lesssim$ 50-dimensional) parameter estimation tasks [2, 55, 65], it has yet to cover higher-dimensionality problems like image reconstruction, which are an essential component in astrophysical and cosmological data analysis.

Here, we focus on the recovery of cosmological initial conditions from late-time density fields. This task is a challenging test case for new algorithms thanks to the

non-linear, non-local mapping from the Gaussian target to the observation. Cosmic inflation predicts the density in the early universe to be highly homogeneous, with tiny density fluctuations that are extremely well described as a Gaussian random field. These density perturbations then gradually grow over cosmic time due to gravity and eventually collapse into the non-Gaussian "Cosmic Web" structure observed today [43]. The reconstruction of the initial density field from late-time observations is an ill-posed problem (the early-to-late mapping is not injective on small scales, [e.g., 319]). Therefore, there is an entire *distribution* of possible initial conditions consistent with a given late-time density field.

Our contribution. We frame the task of field reconstruction (or, viewed from a non-physical point of view, image reconstruction) as a parameter inference problem. We combine the power of SBI in solving parametric inverse problems together with the scalability offered by autoregressive models. Autoregressive models have established their versatility in tackling high-dimensional distribution estimation tasks by breaking down the joint distribution into a product of conditionals [288, 289], and have been successful in conditional image modeling [320]. Additionally, we employ a Gibbs sampling algorithm based on exact data augmentation (GEDA) [321] to efficiently sample image parameter posteriors. We will formulate our method in a generic way to emphasize its applicability to a wide range of field/image reconstruction problems. Importantly, our approach accommodates arbitrary *non-differentiable* forward simulators.

Related work. The problem of inferring cosmological initial conditions has been studied since the late 1980s (see e.g. Refs. [322–327] for classical papers), for instance by applying the least-action principle or using optimal transport. In the last decade, Bayesian models have been formulated for this task [e.g., 44], many of which rely on differentiable forward models [328, 329], in conjunction with Hamiltonian Monte Carlo sampling. Recently, machine learning methods such as convolutional neural networks [330, 331], variational inference [332], recurrent inference machines [333], and score-based modeling [334] have also been explored in this context.

5.2 Methodology

We structure our methodology as follows. First, in Section 5.2.1, we set up the problem in terms of a simple hierarchical field simulator. We then introduce this work's main contribution: autoregressive gaussian likelihood estimation (Section 5.2.2) and high-dimensional gaussian posterior sampling through GEDA (Section 5.2.3).

5.2.1 Problem setup

Let us assume we have a simple hierarchical simulator

$$p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{5.1}$$

where $\boldsymbol{x} \in \mathbb{R}^{N\times N}$ is observed and $\boldsymbol{\theta} \in \mathbb{R}^{N\times N}$ are image parameters (pixel values). Here, $p(\boldsymbol{x}|\boldsymbol{\theta})$ can include non-linear, non-local transformations and non-Gaussian noise, and it is only implicitly defined through a forward simulator. We will discuss how we can, for a given observation $\boldsymbol{x}_o$, estimate an approximate but computationally efficient Gaussian likelihood that locally resembles $p(\boldsymbol{x}_o|\boldsymbol{\theta})$ for a target observation $\boldsymbol{x}_o$. A fast surrogate can then be leveraged for downstream image analysis tasks.

5.2.2 Autoregressive gaussian likelihood estimation

Firstly, in order to obtain locally optimal data summaries $\boldsymbol{s}(\boldsymbol{x})$ for the image reconstruction task, we use NRE to estimate the marginal, pixel-wise ratios[1]

$$\hat{r}(s_i(\boldsymbol{x}); \theta_i) \equiv \frac{p(s_i, \theta_i)}{p(s_i)p(\theta_i)} \; . \tag{5.2}$$

We assume here that both the joint and marginal distributions can be approximated as Gaussians, whereas the mapping $s_i(\boldsymbol{x})$ is an arbitrary learnable function (usually a neural network). Means and covariances are estimated on-the-fly during training (similarly to batch normalization [335]) and are not represented as learnable parameters. Secondly, in order to obtain an estimate of the joint likelihood $p(\boldsymbol{x}|\boldsymbol{\theta})$, we proceed as follows. We split the problem auto-regressively along the observation axis, and we use NRE to estimate

$$\begin{aligned} \frac{p(\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x})} \simeq \frac{p(\boldsymbol{s}(\boldsymbol{x})|\boldsymbol{\theta})}{p(\boldsymbol{s}(\boldsymbol{x}))} &= \prod_{i=1}^{N^2} \frac{p(s_i|\boldsymbol{s}_{1:i-1}, \boldsymbol{\theta})}{p(s_i|\boldsymbol{s}_{1:i-1})} \\ &\simeq \prod_{i=1}^{N^2} \hat{r}(s_i; l_i, \theta_i) = \prod_{i=1}^{N^2} \frac{p(s_i, l_i, \theta_i)}{p(s_i)p(l_i, \theta_i)} \; , \end{aligned} \tag{5.3}$$

where we have introduced $l_i = (\boldsymbol{L}(\boldsymbol{s}))_i$ with $\boldsymbol{L}$ a (generally non-linear) autoregressive function. Again, we assume that the individual functions $p(s_i, l_i, \theta_i)$ are (three-dimensional multivariate) Gaussians.

By rewriting the above components (for a complete derivation and definitions of $\boldsymbol{Q}_{\text{like}}$ and $\boldsymbol{b}$ see Appendix 5.A), we can obtain the likelihood function in the so-called

[1] In this work we use the following notation for ratio estimators $\hat{r}(a; b) = \frac{p(a,b)}{p(a)p(b)} = \frac{p(a|b)}{p(a)}$. If necessary, multiple variables are comma separated, for example $\hat{r}(a; b, c) = \frac{p(a,b,c)}{p(a)p(b,c)}$.

information form,

$$\ln p(\boldsymbol{x}|\boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{\theta}^T \boldsymbol{Q}_{\text{like}} \boldsymbol{\theta} + \boldsymbol{b}\boldsymbol{\theta} + C(\boldsymbol{s}) \; . \tag{5.4}$$

In this paper, we assume that the precision matrix $\boldsymbol{Q}_{\text{like}}$ is diagonal. However, correlations between data summaries $\boldsymbol{s}$ are accounted for through the autoregressive function mentioned above; also, each component $s_i(\boldsymbol{x})$ of the data summary may depend on all components of $\boldsymbol{x}$ and thus accounts for cross-pixel information.

To enhance the robustness of Bayesian approaches in data analysis, frequently likelihood tempering techniques are employed that result in conservative estimates [336, 337]. Tempering the likelihood amounts to raising it to a fractional power $\gamma \in [0, 1]$, leading to progressively coarser posterior samples, with the extreme case of $\gamma = 0$ ($\gamma = 1$) corresponding to samples drawn from the prior (posterior) distribution.

5.2.3 High-dimensional gaussian posterior sampling

In our Bayesian framework and assuming Gaussian distributions, we combine the estimated likelihood with the known prior to derive the posterior in the information form

$$\ln p(\boldsymbol{\theta}|\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{\theta}^T \underbrace{(\boldsymbol{Q}_{\text{like}} + \boldsymbol{Q}_{\text{prior}})}_{\boldsymbol{Q}_{\text{post}}} \boldsymbol{\theta} + \boldsymbol{b}\boldsymbol{\theta} + C'(\boldsymbol{s}) \; , \tag{5.5}$$

where we have assumed a zero-mean prior, consistent with the physical problem we study in Section 5.3. We use the conjugate gradient (CG) algorithm[2] to compute the maximum-a-posteriori (MAP) estimate $\boldsymbol{\theta}_{\text{MAP}}$ of the image parameters $\boldsymbol{\theta}$ by solving the linear system $\boldsymbol{Q}_{\text{post}}\boldsymbol{\theta} = \boldsymbol{b}$. The surrogate Gaussian posterior distribution is then given by $p(\boldsymbol{\theta}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\theta}_{\text{MAP}}, \boldsymbol{Q}_{\text{post}}^{-1})$. Once we have the posterior distribution in the above form, we simply have to sample from it. Various techniques have been presented over time to tackle the problem of efficient sampling from a high-dimensional Gaussian distribution (for a recent review, see Ref. [338]). To obtain the Gaussian posterior samples, we use a Gibbs sampler based on the generalized exact data augmentation algorithm (GEDA) [321]. GEDA solves the problem of high-dimensional Gaussian sampling specifically for distributions whose precision matrix can be expressed as $\boldsymbol{Q} = \boldsymbol{Q}_1 + \boldsymbol{Q}_2$ by exploiting specific properties of $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$. Crucially, $\boldsymbol{Q}_{\text{post}}$ satisfies these constraints.

In general, data augmentation approaches target precision matrices of the form $\boldsymbol{Q} = \boldsymbol{Q}_1 + \boldsymbol{Q}_2$, which naturally arise from the statistical model under investigation. Taking advantage of potential specific structures of $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$, data augmentation strategies introduce one (or several) auxiliary variable $\boldsymbol{u} \in \mathbb{R}^d$ such that the joint

[2] We use a slightly modified implementation of the preconditioned CG algorithm from https://github.com/sbarratt/torch_cg.

distribution of the pair $(\boldsymbol{u}, \boldsymbol{\theta})$ yields simple conditional distributions, thus sampling steps for a Gibbs sampler. One can recover the target distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{Q}^{-1})$ via marginalization of the auxiliary variable $\boldsymbol{u}$, either exactly (as in exact data augmentation schemes, like GEDA) or in an asymptotic regime.

In GEDA, as described in Ref. [321], the underlying assumption is that the precision matrix $\boldsymbol{Q}$ can be split as follows:

$$\begin{aligned} \boldsymbol{Q} &= \boldsymbol{Q}_1 + \boldsymbol{Q}_2 \\ \boldsymbol{Q}_1 &= \boldsymbol{G}_1^T \boldsymbol{D}_1 \boldsymbol{G}_1 \\ \boldsymbol{Q}_2 &= \boldsymbol{U}_2^T \boldsymbol{D}_2 \boldsymbol{U}_2 \,, \end{aligned} \tag{5.6}$$

where $\boldsymbol{G}_1$ is arbitrary, $\boldsymbol{D}_1$ is diagonal and positive definite, $\boldsymbol{U}_2$ is unitary, and $\boldsymbol{D}_2$ is diagonal. GEDA introduces two auxiliary variables, $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$, such that the joint distribution is

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{u}_1, \boldsymbol{u}_2) \propto \exp\left(-\frac{1}{2} \left[(\boldsymbol{\theta} - \boldsymbol{\mu})^T \boldsymbol{Q} (\boldsymbol{\theta} - \boldsymbol{\mu}) + (\boldsymbol{u}_1 - \boldsymbol{\theta})^T \boldsymbol{R} (\boldsymbol{u}_1 - \boldsymbol{\theta}) \right] \right) \\ \times \exp\left(-\frac{1}{2} (\boldsymbol{u}_2 - \boldsymbol{G}_1 \boldsymbol{u}_1)^T \boldsymbol{D}_1 (\boldsymbol{u}_2 - \boldsymbol{G}_1 \boldsymbol{u}_1) \right) \,, \end{aligned} \tag{5.7}$$

where $\boldsymbol{R} = \omega^{-1} \boldsymbol{I}_d - \boldsymbol{Q}_1$, and ω is a positive hyper-parameter of the algorithm that must obey $1 < \omega < 1/|\boldsymbol{Q}_1|$, meaning that $1/\omega$ should be larger than the largest singular value of $\boldsymbol{Q}_1$.

This joint distribution yields conditional Gaussian distributions with diagonal covariance matrices for both $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ that can be sampled efficiently by a Gibbs sampler with the following steps:

$$\begin{aligned} &1. \quad \boldsymbol{u}_2 \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{u}_2}, \boldsymbol{Q}_{\boldsymbol{u}_2}^{-1}) \quad \text{with } \boldsymbol{\mu}_{\boldsymbol{u}_2} = \boldsymbol{G}_1 \boldsymbol{u}_1 \text{ and } \boldsymbol{Q}_{\boldsymbol{u}_2} = \boldsymbol{D}_1, \\ &2. \quad \boldsymbol{u}_1 \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{u}_1}, \boldsymbol{Q}_{\boldsymbol{u}_1}^{-1}) \quad \text{with } \boldsymbol{\mu}_{\boldsymbol{u}_1} = \boldsymbol{\theta} - \omega(\boldsymbol{Q}_1 \boldsymbol{\theta} - \boldsymbol{G}_1^T \boldsymbol{D}_1^{-1} \boldsymbol{u}_2) \text{ and } \boldsymbol{Q}_{\boldsymbol{u}_1} = \omega^{-1} \boldsymbol{I}_d, \\ &3. \quad \boldsymbol{u}_1 \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{u}_1}, \boldsymbol{Q}_{\boldsymbol{u}_1}^{-1}) \quad \text{with } \boldsymbol{\mu} = \boldsymbol{Q}(\boldsymbol{R}\boldsymbol{u}_1 + \boldsymbol{Q}\boldsymbol{\mu}) \text{ and } \boldsymbol{Q} = \omega^{-1} \boldsymbol{I}_d + \boldsymbol{Q}_2 \,. \end{aligned} \tag{5.8}$$

In our setup the dimensionality is equivalent to the number of pixels $d = N^2$, the covariance matrix is $\boldsymbol{Q} = \boldsymbol{Q}_{\text{post}} = \boldsymbol{Q}_{\text{like}} + \boldsymbol{Q}_{\text{prior}}$, and $\boldsymbol{G}_1 = \boldsymbol{I}$.

5.3 Experiment

To demonstrate the efficiency of our method, we apply it to the task of reconstructing the initial conditions of the universe.

Forward model. In this proof-of-concept study, we consider the two-dimensional case and assume Einstein–de Sitter cosmology (i.e. non-relativistic, collisionless matter

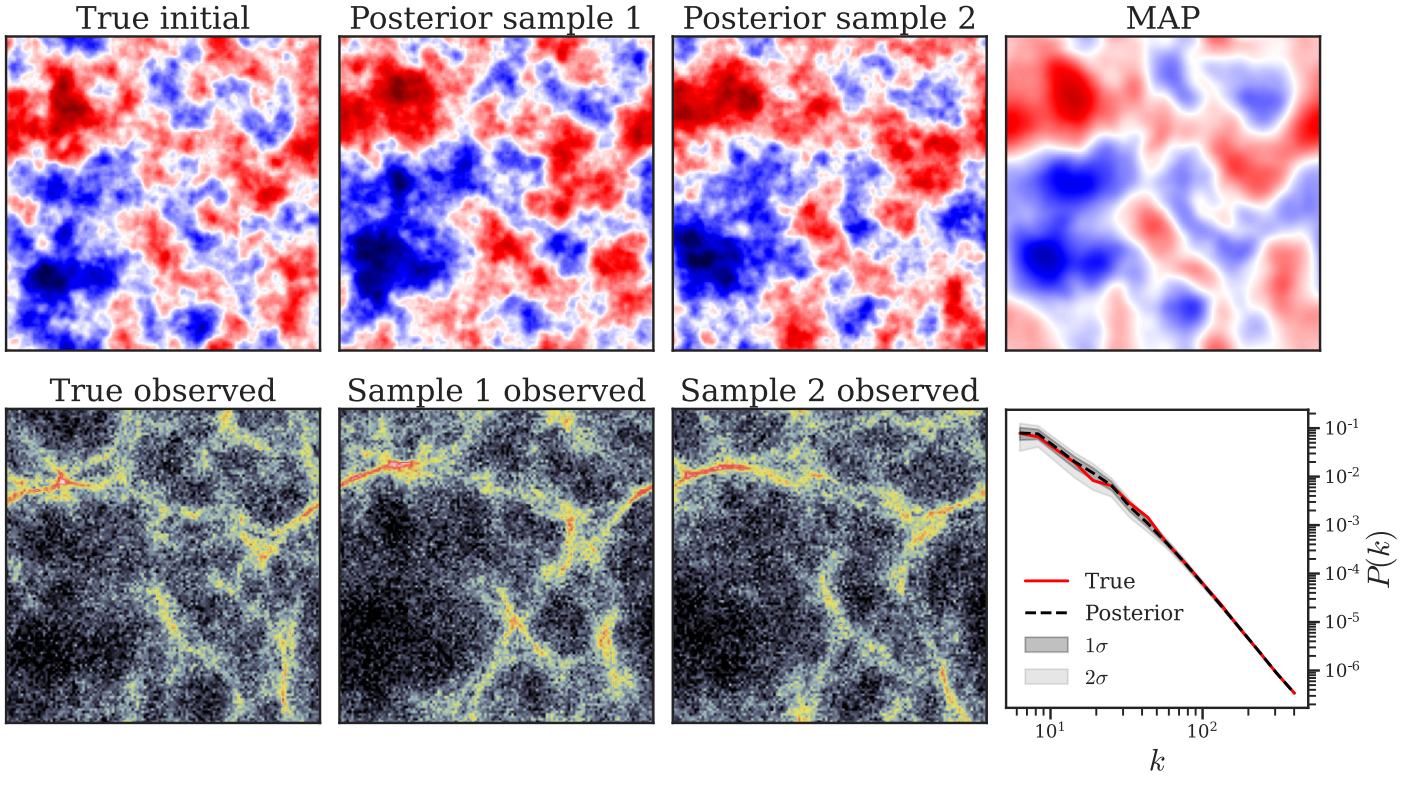

Figure 5.1: Reconstruction of cosmological initial conditions in 2D. *Top left:* True initial density $\boldsymbol{\theta}_o$. *Bottom left*: True observation $\boldsymbol{x}_o$, i.e. the (logarithm of the) late-time density evolved from the initial conditions $\boldsymbol{\theta}_o$, corrupted by uncorrelated Gaussian noise. *Top center:* Two samples drawn from the posterior $p(\boldsymbol{\theta}|\boldsymbol{x}_o)$. *Bottom center:* Observations computed from the posterior samples shown above. *Upper right:* Maximum-a-posteriori probability (MAP) estimate $\boldsymbol{\theta}_{\mathrm{MAP}}$ of the initial conditions. *Bottom right:* Distribution of the reconstructed initial power spectrum.

only). While our framework readily supports marginalizing over image parameters such as the power spectra of the target fields, we use a fixed power-law power spectrum for the initial density contrast $\boldsymbol{\theta} = \boldsymbol{\delta}_{\text{ini}} \in \mathbb{R}^{128\times128}$, which is the target of our inference. As a forward model, we use second-order Lagrangian perturbation theory (2LPT, see Refs. [339, 340] and Appendix 5.B) and evolve the initial density to a time when non-linear structures have formed. The observation is then given by $\boldsymbol{x} = \log_{10}[1.1 + \boldsymbol{\delta}_{\text{final}}] + \boldsymbol{\varepsilon}$, where $\boldsymbol{\delta}_{\text{final}}$ is the density contrast at final time, the logarithm is applied element-wise, and we add $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ with $\sigma = 0.15$ as a simplistic model for observational noise.

Training strategy. To obtain pixel-wise summaries $\boldsymbol{s}(\boldsymbol{x})$, we use a standard U-Net [341]. In our current implementation, we use an autoregressive convolution [e.g., 320] as the autoregressive function $\boldsymbol{L}$. In our experiments, we observed that overly small or large kernel sizes result in somewhat biased posteriors. Interestingly, we obtain the best results with a physically motivated kernel size for the convolution, i.e. when choosing it in such as way that the "radius of influence" of the autoregressive convolution (i.e. roughly half the kernel size) matches the typical distance that particles (or fluid elements) have traveled by the end time – a quantity known as displacement. Specifically, the mean displacement in our case is ~ 5 pixels, so we take the kernel size to be 11×11 pixels. While we find this choice to be optimal, slight bias still occurs occasionally, potentially due to the spatial heterogeneity of the displacements. Therefore, to be conservative, we use a tempered likelihood with $\gamma = 0.5$, trading some of the statistical constraining power of our method for increased unbiasedness. The importance of the specific choice of $\boldsymbol{L}$ (and in particular its radius of influence) indicates that a locally adaptive or multi-scale approach for $\boldsymbol{L}$ is a promising avenue for future exploration. We train our model on 1000 $(\boldsymbol{x}, \boldsymbol{\theta})$-pairs for 15 epochs, which only takes ~ 3 minutes on a laptop GPU for this experiment. Obtaining the MAP estimate with the CG algorithm takes less than 3 seconds. To obtain posterior samples with GEDA, we perform 300 sampling steps, which takes < 1 second per sample (and multiple samples can be drawn in batches).

Results. The left column of Figure 5.1 shows a target initial density contrast $\boldsymbol{\theta}_o$ (*top*), together with a resulting noisy observation at late time $\boldsymbol{x}_o$ (*bottom*). In the two center panels, we plot samples drawn from the (tempered) posterior $p(\boldsymbol{\theta}|\boldsymbol{x}_o)$, together with an observation of each sample. The upper right panel shows the MAP estimate $\boldsymbol{\theta}_{\text{MAP}}$. The initial density field is faithfully reconstructed on large scales whereas, as expected, small-scale information remains unconstrained. Finally, the lower right panel shows that the power spectra of the posterior samples are consistent with the target field. The excellent agreement between reconstructed and true power spectra on small scales is aided by the fact that the power spectrum (which is fixed in our

example) directly enters the GEDA sampling (see Equation 5.6 in Section 5.2.3, where the prior precision matrix $\boldsymbol{Q}_2$ is diagonalized by the Fourier transform, with the power spectrum on the diagonal of $\boldsymbol{D}_2$). We will present a detailed quantitative analysis of the results obtained with our framework in a separate publication.

5.4 Discussion and conclusions

We have introduced a framework for Bayesian field/image reconstruction by combining SBI, autoregressive modeling, and a Gibbs sampling algorithm based on exact data augmentation (GEDA). We presented promising results for a toy example related to reconstructing the initial conditions of the universe. In view of its *remarkable speed, low simulation costs, and the fact that it works for general non-differentiable simulators*, we expect our method to be capable of handling significantly higher-dimensional problems, including 3D cosmological simulations.

There are multiple promising avenues to be explored in future work. For many applications—such as the reconstruction of cosmological initial conditions—formulating the problem in Fourier space can be expected to produce significantly tighter posteriors, as the input-to-output mapping is exactly invertible on large to intermediate scales, and the stochasticity of the reconstruction only affects small scales. Alternatively, wavelets [e.g., 342] or other multiscale techniques could be exploited. In this context (but also more generally), other choices for the autoregressive function $\boldsymbol{L}$ are worthwhile exploring, which could, for instance, implement the hierarchy between the different scales. Another promising enhancement involves harnessing the sequential aspect of SBI techniques. In principle, a viable strategy is to employ an adaptive scheduler to control the value of γ to draw new training samples for sequential inference rounds. Moreover, being SBI-based, our proposed method is capable of marginalizing over cosmological parameters or, alternatively, infer them in addition to the phases of the initial random field.

Let us also comment on the current limitations of our method: while the assumption of Gaussianity for the prior is justified for reconstructing cosmological initial conditions, using a Gaussian likelihood is an approximation whose justification depends on the specific task at hand. We take cross-pixel information into account through the summary statistics and an autoregressive function, but we currently model the precision matrix of the likelihood as being diagonal. In addition, the susceptibility of our framework to issues encountered in the context of autoregressive models such as exposure bias [343] merits further investigation.

Finally, we remark that the forward model generating "observations" $\boldsymbol{x}$ does not necessarily need to be a physical one, and our framework also holds great promise for generic image reconstruction problems consisting in inferring one image from another.

Appendix

5.A Derivation of the likelihood in the information form

In this appendix we show how to derive the quadratic term $\boldsymbol{Q}_{\text{like}}$ and the linear term $\boldsymbol{b}$ of the likelihood in the information form (Equation 5.4) starting from Equation 5.3. Having approximated $p(s_i, l_i, \theta_i)$ with a 3-dimensional Gaussian distribution, we can write the log-likelihood as

$$\ln p(\boldsymbol{s}(\boldsymbol{x})|\boldsymbol{\theta}) = \sum_{i=1}^{N^2} \ln \frac{p(s_i, l_i, \theta_i)}{p(l_i, \theta_i)} = \sum_{i=1}^{N^2} \ln \frac{\mathcal{N}(\mu_{(s_i,l_i,\theta_i)}, \Sigma_{(s_i,l_i,\theta_i)})}{\mathcal{N}(\mu_{(l_i,\theta_i)}, \Sigma_{(l_i,\theta_i)})} \,. \tag{5.9}$$

Let us consider just one element i and drop the index i for simplicity of notation. Expanding the above expression for one component, we obtain

$$-\frac{1}{2}\begin{pmatrix} s-\mu_s \\ l-\mu_l \\ \theta-\mu_\theta \end{pmatrix}^T \underbrace{\begin{pmatrix} \Sigma_{ss} & \Sigma_{sl} & \Sigma_{s\theta} \\ \Sigma_{ls} & \Sigma_{ll} & \Sigma_{l\theta} \\ \Sigma_{\theta s} & \Sigma_{\theta l} & \Sigma_{\theta\theta} \end{pmatrix}^{-1}}_{\equiv Q' = \begin{pmatrix} Q'_{ss} & Q'_{sl} & Q'_{s\theta} \\ Q'_{ls} & Q'_{ll} & Q'_{l\theta} \\ Q'_{\theta s} & Q'_{\theta l} & Q'_{\theta\theta} \end{pmatrix}} \begin{pmatrix} s-\mu_s \\ l-\mu_l \\ \theta-\mu_\theta \end{pmatrix} + \frac{1}{2}\begin{pmatrix} l-\mu_l \\ \theta-\mu_\theta \end{pmatrix}^T \underbrace{\begin{pmatrix} \Sigma_{ll} & \Sigma_{l\theta} \\ \Sigma_{\theta l} & \Sigma_{\theta\theta} \end{pmatrix}^{-1}}_{\equiv Q'' = \begin{pmatrix} Q''_{ll} & Q''_{l\theta} \\ Q''_{\theta l} & Q''_{\theta\theta} \end{pmatrix}} \begin{pmatrix} l-\mu_l \\ \theta-\mu_\theta \end{pmatrix} . \tag{5.10}$$

For a specific observation $\boldsymbol{x}$ we can compute $\boldsymbol{s}$ and therefore $\boldsymbol{l}$, hence individual components s and l (where we have dropped the index i). We can then read off the expression in Equation 5.10 the terms that depend on θ. We group the rest into an irrelevant constant $C(s)$. In this way, the *quadratic term* in θ in Equation 5.10 is

$$-\frac{1}{2} Q_{\theta\theta}\theta^2 \quad \text{with} \quad Q_{\theta\theta} \equiv Q'_{\theta\theta} - Q''_{\theta\theta} \,.$$

Each $Q_{\theta\theta}$ is a diagonal entry of $\boldsymbol{Q}_{\text{like}}$ in Equation 5.4, while non-diagonal components are 0 by construction. In principle, correlations between pixels in the data $\boldsymbol{x}$ can be modeled and we plan to explore it in future work. The *linear term* in θ in Equation 5.10 is

$$b \equiv [-(s-\mu_s)Q'_{s\theta} - (l-\mu_l)Q_{l\theta} + \mu_\theta Q_{\theta\theta}] \,. \tag{5.11}$$

Each component b composes the likelihood linear term $\boldsymbol{b}$.

5.B Second-order Lagrangian perturbation theory (2LPT)

Second-order Lagrangian perturbation theory (2LPT) describes the cosmological dynamics of a cold, collisionless medium. It is based on a fluid description of the Vlasov–Poisson system [e.g. 326] and therefore ceases to be valid as soon as particle trajectories cross (so-called "shell-crossing"). Although the density fields we consider herein are well in the post-shell-crossing regime, 2LPT still serves as a suitable forward model for validating our image reconstruction method.

The central quantity in 2LPT is the displacement $\boldsymbol{\psi}(\boldsymbol{q}, a) = \boldsymbol{p}(\boldsymbol{q}, a) - \boldsymbol{q}$, i.e. the vector pointing from each Lagrangian ("initial") position $\boldsymbol{q}$ to the associated Eulerian position $\boldsymbol{p}(\boldsymbol{q}, a)$ at scale-factor time a on the characteristic curve originating at $\boldsymbol{q}$. LPT expands the displacement as a Taylor series w.r.t. a, with a purely space-dependent coefficient $\boldsymbol{\psi}^{(n)}(\boldsymbol{q})$ at the n-th order. For 2LPT, this yields

$$\boldsymbol{\psi}(\boldsymbol{q}, a) = a\boldsymbol{\psi}^{(1)}(\boldsymbol{q}) + a^2\boldsymbol{\psi}^{(2)}(\boldsymbol{q}) \,, \tag{5.12}$$

where

$$\boldsymbol{\psi}^{(1)} = -\boldsymbol{\nabla}_{\boldsymbol{q}}\phi^0 \,, \tag{5.13a}$$

$$\boldsymbol{\psi}^{(2)} = -\frac{3}{7}\boldsymbol{\nabla}_{\boldsymbol{q}}\,\Delta_{\boldsymbol{q}}^{-1}\left[(\partial_{q_1,q_1}\phi^0)\,(\partial_{q_2,q_2}\phi^0) - (\partial_{q_1,q_2}\phi^0)^2\right] \,. \tag{5.13b}$$

Here, ϕ^0 is the initial gravitational potential, and the scale factor a simply acts as a time variable and has no physical meaning in 2D. Note that we consider Einstein–de Sitter cosmology ($\Omega_m = 1$); for extensions to the ΛCDM case and higher-order LPT, see e.g. Ref. [344] and references therein.

Given the 2LPT displacement $\boldsymbol{\psi}$, the density contrast δ is given by

$$1 + \delta(\boldsymbol{p}(\boldsymbol{q}, a)) = \int \delta_{\mathrm{D}}(\boldsymbol{p}(\boldsymbol{q}, a) - \boldsymbol{q} - \boldsymbol{\psi}(\boldsymbol{q}, a))\,\mathrm{d}^2q \,, \tag{5.14}$$

where δ_{D} is the Dirac δ-distribution. In practice, we use cloud-in-cell interpolation [e.g. 345] to compute the density contrast on a Cartesian grid. Finally, let us remark that since we view the discrete density contrast as an $N \times N$-dimensional image in the main part of this work, we use the bold symbol $\boldsymbol{\delta}$ in that context.

6 Simulation-based inference for point sources

Statistical inference of population parameters of astrophysical sources is challenging. It requires accounting for selection effects, which stem from the artificial separation between bright detected and dim undetected sources that is introduced by the analysis pipeline itself. In this chapter, we show that these effects can be modeled self-consistently in the context of sequential SBI. Our approach couples source detection and catalog-based inference in a principled framework that derives from the TMNRE algorithm. It relies on the realization that detection can be interpreted as prior truncation. We outline the algorithm, and show first promising results.

This chapter is based on work from [4].

6.1 Introduction

Point sources detection is crucial for astronomical surveys, and is the cornerstone for the compilation of source catalogues. Those source catalogues are then typically the basis for the inference of physical parameters that describe the sources at the population level. Upcoming astronomical facilities, such as the Square Kilometer Array (SKA) [346] and the Cherenkov Telescope Array (CTA) [347] will deliver large and complex datasets. In order to leverage their full potential, it is urgent to develop robust and automated source detection and source population parameters inference algorithms.

Machine learning is a particularly promising tool to address this big data challenge. Recent developments in deep learning and more generally automatic differentiation frameworks [54] are increasingly used for tackling difficult astronomical data analysis with the goal to extricate more information from the available data.

The capability of deep learning techniques of point sources detection and population characterization has been demonstrated across different wavelengths surveys, e.g. in γ-ray data [87, 348–350], radio data [351–354] and cosmic microwave background data [355]. In particular simulation-based machine learning approaches can be highly flexible, allowing to tailor developed pipelines to specific telescopes and science cases. A range of SBI algorithms have been proposed in the literature (see Ref. [52] for a review). An appealing feature is that they generally allow to directly estimate marginal posteriors for parameters of interest [120]. Furthermore, *sequential* SBI approaches [146, 149, 150] have been shown to be particularly simulation efficient. Among those, TMNRE [120, 160] is a sequential SBI approach based on NRE [153], which particularly well composes with marginalization.

Our contribution. Here, we present a strategy for how to use TMNRE to simultaneously perform source detection and population-level parameters inference. This enables to self-consistently combine information from both detected and sub-threshold sources, without being affected by detection biases. The key idea is to recast the traditional concept of *source detection* in terms of *prior truncation*. This will allow us to *distill* information of bright sources directly into the simulation model. Our proposed method is highly interpretable since it resembles components of traditional survey analysis workflows.

6.2 Methodology

We structure our methodology as follows. First, we introduce the problem setup, phrasing simulating point source maps in terms of a Bayesian hierarchical forward model in Section 6.2.1. Secondly, in Section 6.2.2, we perform point sources detection and estimate the point sources sensitivity function of our method. We then show how to include our catalogue of detected point sources in the forward model, recasting source detection as prior truncation, in Section 6.2.3. Lastly, we illustrate how to self-consistently perform point sources population parameters inference, by combining information from both detected and sub-threshold sources, in Section 6.2.4. A schematic overview of the inference framework is shown in Figure 6.1.

6.2.1 Problem setup

We consider here a simple Bayesian hierarchical source model,

$$p(\boldsymbol{x}, \vec{\boldsymbol{s}}, \boldsymbol{\theta}) = p(\boldsymbol{x} \mid \vec{\boldsymbol{s}}) p(\boldsymbol{\theta}) \prod_{i=1}^{N} p(\boldsymbol{s}_i \mid \boldsymbol{\theta}) \,, \tag{6.1}$$

Parameter		Prior
Population parameters		
number of point sources	N	$\mathcal{U}(10, 500)$
flux distribution parameter	Σ	$\mathcal{U}(1, 3)$
spatial distribution parameter	h	$\mathcal{U}(1, 20)$
Point source parameters		
flux	F	$\log \mathcal{N}(1, \Sigma)$
position	$\Omega \equiv (l, b)$	$(\mathcal{N}(0, 20), \mathcal{N}(0, h))$

Table 1: Point source simulation model parameters and priors.

where $\boldsymbol{x}$ is the observed sky map, $\boldsymbol{s}_i \equiv (F_i, \Omega_i)$ denotes the flux F_i and position Ω_i of point source i, and $\boldsymbol{\theta} \equiv \{N, \Sigma, h\}$ collects source population parameters, namely the number of sources N, and the parameters Σ and h that control the flux and spatial distributions respectively. To generate an observation, first, we sample point sources population parameters $\boldsymbol{\theta} \equiv \{N, \Sigma, h\}$ from their priors, given in Table 1. Then, for each point source, we draw its flux F and position on the map $\Omega \equiv (l, b)$ from the priors given in Table 1. We then generate a 128×128 pixels map. To model instrumental effects, we add a PSF with Gaussian kernel standard deviation $\epsilon = 1.5$ and Poisson noise, obtaining the final simulated map $\boldsymbol{x}$. We show examples of our simulated maps in the top row of the right panel of Figure 6.2.

6.2.2 Source detection

For source detection we consider the following likelihood-to-evidence ratio,[1]

$$r_1(\Omega, F_{th}; \boldsymbol{x}) \equiv \frac{p(\mathbb{I}_{\boldsymbol{x}}(F \geq F_{th}) = 1, \Omega \mid \boldsymbol{x})}{p(\mathbb{I}_{\boldsymbol{x}}(F \geq F_{th}) = 1, \Omega)} \,. \tag{6.2}$$

Here, the denominator corresponds to the prior probability of having a source at position Ω with a flux $F \geq F_{th}$ that exceeds some threshold flux F_{th}. The numerator is the corresponding posterior. We model the source detection ratio estimator in Equation (6.2) as an image-to-image neural network that solves a binary classification problem in each image pixel. For simulated data, we call a simulated source $\boldsymbol{s}_i$ 'detected' when there is a corresponding compact region as function of Ω where the detection significance is above threshold, $r_1(\Omega, F_{th}; \boldsymbol{x}) > 5$. This effectively leads to a

[1] In the following sections we will use the notation $r(a; b \mid c) = \frac{p(a,b|c)}{p(a|c)p(b|c)}$, where with '|' we refer to conditioning on specific variables for all factors in the ratio definition. If necessary, multiple variables are comma separated, for example $r(a, b; c \mid d) = \frac{p(a,b,c|d)}{p(a,b|d)p(c|d)}$. Training conditional ratios is a straight-forward extension of NRE, as seen in Section 4.2.1.

split between sources that are clearly identifiable and 'sub-threshold' sources that are difficult to detect as individual instances. Below, we assign the detection label $d_i = 1$ ($d_i = 0$) to detected (undetected) sources.

In order to characterize the split between detected and sub-threshold sources, we introduce a source-sensitivity function, $S(F, \Omega)$, which provides the probability that a source with flux F and at position Ω would be detected by the ratio estimator in Equation (6.2). This function can be estimated by training the ratio estimator

$$r_2(d; F, \Omega, \boldsymbol{x}) \equiv \frac{p(d \mid F, \Omega, \boldsymbol{x})}{p(d)} \tag{6.3}$$

which is marginalised over all other sources and source parameters. By omitting the dependence on the map, $\boldsymbol{x}$, which is then effectively marginalized, the ratio estimator can then simply be modeled as a $\mathbb{R}^3 \to \mathbb{R}$ MLP. The source-sensitivity function can then be estimated as

$$S(F, \Omega) = \sigma\left(\log\left(\frac{p(d = 1 \mid F, \Omega)}{p(d = 0 \mid F, \Omega)}\right)\right), \tag{6.4}$$

where we have introduced the sigmoid function $\sigma(y) \equiv 1/(1 + e^{-y})$.

Once we know the sensitivity function $S(F, \Omega)$, we can make the concept of source detection part of our model as follows. In a random realization, each source i will be either detected, $d_i = 1$ (with probability $S(F_i, \Omega_i)$), or not detected, $d_i = 0$ (with probability $1 - S(F_i, \Omega_i)$). To keep notation simple, we omit d_i and instead group detected and non-detected (sub-treshold) sources together in vectors with corresponding subscripts, and write $\vec{\boldsymbol{s}} \equiv (\vec{\boldsymbol{s}}_{det}, \vec{\boldsymbol{s}}_{sub})$. Taking into consideration this split, our model becomes

$$p(\boldsymbol{x}, \vec{\boldsymbol{s}}_{det}, \vec{\boldsymbol{s}}_{sub}, \boldsymbol{\theta}) = p(\boldsymbol{x} \mid \vec{\boldsymbol{s}}_{det}, \vec{\boldsymbol{s}}_{sub}) p(\vec{\boldsymbol{s}}_{det} \mid \boldsymbol{\theta}) p(\vec{\boldsymbol{s}}_{sub} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, . \tag{6.5}$$

Importantly, although $p(\vec{\boldsymbol{s}}_{det/sub} \mid \boldsymbol{\theta})$ depends on the sensitivity function $S(F_i, \Omega_i)$, the distribution of all sources $p(\vec{\boldsymbol{s}} \mid \boldsymbol{\theta})$ is the same as in Equation (6.1).

6.2.3 Detection as truncation

Finally, we introduce our truncation scheme. Truncating source priors in Equation (6.1) is difficult due to the the label switching problem ($\boldsymbol{x}$ is invariant under relabeling sources). However, once specific sources are detected, they can be labeled and ordered arbitrarily. Let us assume that N_{det} sources were detected by the ratio estimator in Equation (6.2) in the regions $\mathcal{R}_i$ with $i = 1, \ldots, N_{det}$ for a given observation of reference $\boldsymbol{x}_o$. Our ansatz for the indicator function, which selects a specific prior region, is

$$\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = \prod_{i=1}^{N_{det}} \mathbb{I}_{\boldsymbol{x}_o}(\Omega_i \in \mathcal{R}_i) \mathbb{I}_{\boldsymbol{x}_o}(F_i \geq F_{th}) \, . \tag{6.6}$$

Our truncation strategy is now to focus on the parameter space where $\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1$ for our data of interest $\boldsymbol{x}_o$, in order to reduce training data variance. We can then write our truncated model as

$$\begin{aligned} &p(\boldsymbol{x},\vec{\boldsymbol{s}}_{det},\vec{\boldsymbol{s}}_{sub},\boldsymbol{\theta},\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det})) \\ &\quad = p(\boldsymbol{x} \mid \vec{\boldsymbol{s}}_{det},\vec{\boldsymbol{s}}_{sub})p(\vec{\boldsymbol{s}}_{det} \mid \boldsymbol{\theta},\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}))p(\vec{\boldsymbol{s}}_{sub} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})p(\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) \mid \boldsymbol{\theta}) \,, \end{aligned} \tag{6.7}$$

where detected sources $p(\vec{\boldsymbol{s}}_{det})$ can be sampled only in the relevant parameter space $\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1$.

6.2.4 Population parameters inference with truncation

Given some observation $\boldsymbol{x}_o$, we want to estimate the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x}_o)$. Since the truncation affects parameters $\vec{\boldsymbol{s}}_{det}$ whose prior depends on population parameters $\boldsymbol{\theta}$, the truncation volume is not a constant factor, and the procedure requires extra care. We can estimate the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ by considering the ratio

$$\begin{aligned} r(\boldsymbol{\theta};\boldsymbol{x}) = \frac{p(\boldsymbol{\theta} \mid \boldsymbol{x})}{p(\boldsymbol{\theta})} &\simeq \frac{p(\boldsymbol{\theta} \mid \boldsymbol{x},\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1)}{p(\boldsymbol{\theta})} \\ &= \underbrace{\frac{p(\boldsymbol{\theta} \mid \boldsymbol{x},\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1)}{p(\boldsymbol{\theta} \mid \mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1)}}_{\equiv r_3(\boldsymbol{\theta};\boldsymbol{x}\mid\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det})=1)} \underbrace{\frac{p(\boldsymbol{\theta} \mid \mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1)}{p(\boldsymbol{\theta})}}_{\equiv r(\boldsymbol{\theta};\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det})=1)} \,. \end{aligned} \tag{6.8}$$

The second step in Equation (6.8) corresponds to the truncation approximation. It is exact (in the sense of leaving $p(\boldsymbol{\theta} \mid \boldsymbol{x}_o)$ unaffected) in the limit where $p(\boldsymbol{x}_o \mid \mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 0) \to 0$. It exploits the fact that $\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1$ does not add information beyond what is already known from $\boldsymbol{x}_o$. Training that ratio estimator directly with TMNRE would be challenging since $\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1$ has very small support in the training data. Instead, we split the ratio into two computationally feasible ratios (this is in spirit similar to the telescoping ratio estimation approach presented in Ref. [286]).

The first ratio, $r_3(\boldsymbol{\theta};\boldsymbol{x} \mid \mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1)$, can be estimated by training a peak-count network [356–358] on targeted data, that is truncated to $\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1$. The second ratio can be estimated as

$$\begin{aligned} r(\boldsymbol{\theta};\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1) &= \frac{p(\boldsymbol{\theta} \mid \mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1)}{p(\boldsymbol{\theta})} = \frac{p(\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1 \mid \boldsymbol{\theta})}{p(\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1)} \\ &= \int d\vec{\boldsymbol{s}}_{det} \frac{p(\vec{\boldsymbol{s}}_{det} \mid \boldsymbol{\theta})}{p(\vec{\boldsymbol{s}}_{det})} \frac{p(\vec{\boldsymbol{s}}_{det})\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det})}{p(\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1)} \\ &= \int d\vec{\boldsymbol{s}}_{det} \underbrace{\frac{p(\vec{\boldsymbol{s}}_{det} \mid \boldsymbol{\theta})}{p(\vec{\boldsymbol{s}}_{det})}}_{\equiv r_4(\vec{\boldsymbol{s}}_{det};\boldsymbol{\theta})} p(\vec{\boldsymbol{s}}_{det} \mid \mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1) \,, \end{aligned} \tag{6.9}$$

so it can be estimated by training a network on detected sources lists on untruncated training data. In practice, we generate weighted samples from the full posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ by sampling $\boldsymbol{\theta}, \vec{\boldsymbol{s}}_{det} \sim p(\boldsymbol{\theta})p(\vec{\boldsymbol{s}}_{det} \mid \mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1)$ with weights $w = r_3(\boldsymbol{\theta}; \boldsymbol{x} \mid \mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1) \cdot r(\boldsymbol{\theta}; \mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1)$.

In the process, we trained four ratio estimation networks that are directly connected with traditional source analysis pipeline components: $r_1(\Omega, F_{th}; \boldsymbol{x})$ performs source detection; $r_2(d; F, \Omega, \boldsymbol{x})$ is the source sensitivity function; $r_3(\boldsymbol{\theta}; \boldsymbol{x} \mid \mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1)$ constraints $\boldsymbol{\theta}$ based on sub-threshold sources (because detected sources are assumed to be fixed in the parameter space where $\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1$); and $r_4(\vec{\boldsymbol{s}}_{det}; \boldsymbol{\theta})$ constraints $\boldsymbol{\theta}$ using the detected sources catalog. We show a schematic overview of the inference framework used in this work in Figure 6.1.

6.3 Experiment

We apply the proposed methodology to the simulated target observation $\boldsymbol{x}_o$ shown in Figure 6.2.

Training strategy. We first train the source detection ratio estimator in Equation (6.2) on data simulated from the full model shown in Equation (6.1), and then apply it to $\boldsymbol{x}_o$ to obtain a detection map. From the detection map we derive a catalog of detected sources $\vec{\boldsymbol{s}}_{det}$, and define a truncated parameter space of interest, where $\mathbb{I}_{\boldsymbol{x}_o}(\vec{\boldsymbol{s}}_{det}) = 1$. In order to make source detection part of our model, we train the sensitivity ratio estimator in Equation (6.3) to estimate the sensitivity function $S(F, \Omega)$, illustrated in Figure 6.3. We then generate targeted training data from our truncated simulation model in Equation (6.7). We show samples from the full model and the truncated one in Figure 6.2. Finally, we train two inference networks to capture information regarding population parameters from sub-threshold and detected sources, as explained in Section 6.2. The four neural networks were trained on a NVIDIA GeForce RTX 3080 Ti GPU, the total computational time cost to obtain the results shown in Figure 6.4 is ~ 2 hours.

Results. We show the constraints on population parameters $\boldsymbol{\theta}$ from sub-threshold sources, detected ones, and their combination in Figure 6.4. The different posteriors are consistent with each other, indicating that the proposed inference framework automatically accounts for detection biases. We see that different constraints are dominated by different neural networks, e.g. the number N of point source is better constrained by sub-threshold ones, whereas the spatial distribution parameter h by detected ones. The weaker constraint on the flux parameter Σ inferred from detected sources is due to the fact that in Equation (6.9) we average over different detected

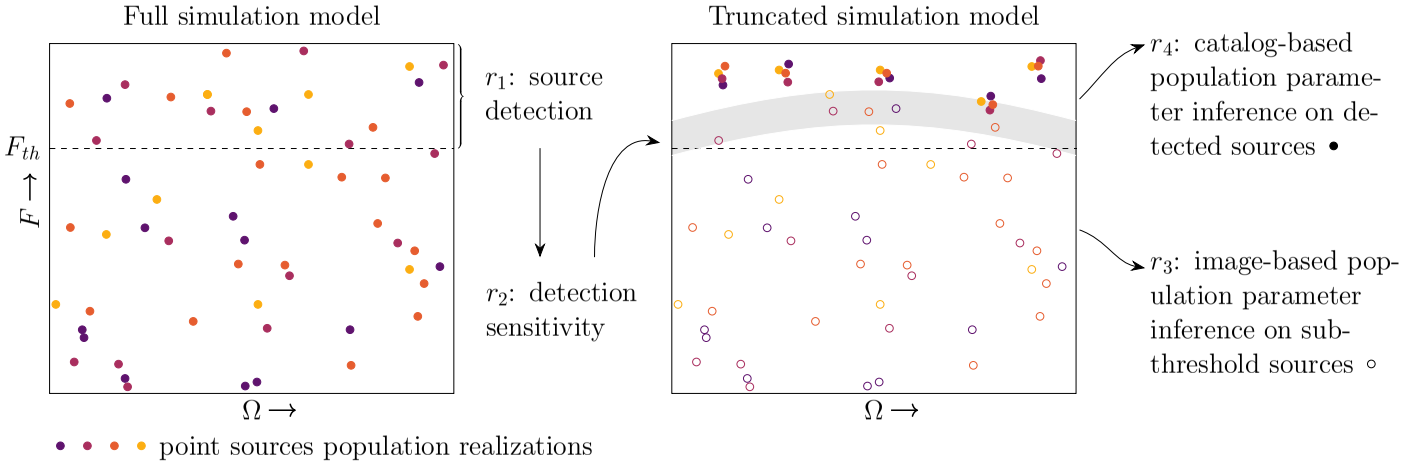

Figure 6.1: Illustration of our inference framework (see Section 6.2 for details). *Left panel*: A source detection network r_1 and the corresponding sensitivity network r_2 are trained based on the full simulation model. *Right panel*: Bright sources are constrained in the truncated simulation model, while sub-threshold sources vary freely. Two inference networks are trained to capture information from sub-threshold sources (r_3) and detected sources (r_4).

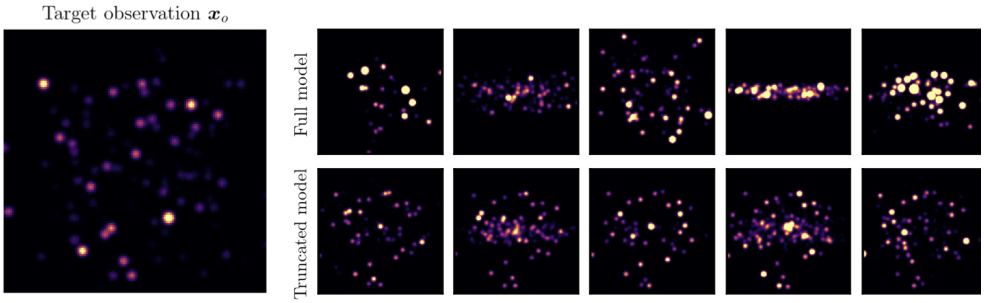

Figure 6.2: *Left:* Observation of reference $\boldsymbol{x}_o$. *Right:* In the first row we show samples from our full simulation model. In the second row we show targeted samples from the truncated model. Targeted data are visually more close to $\boldsymbol{x}_o$, the main dissimilarities are due to different sub-threshold sources and instrumental effects realizations.

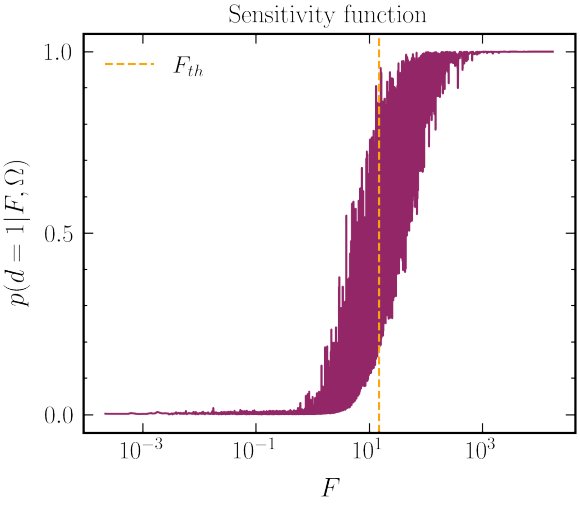

Figure 6.3: Point source sensitivity function $S(F, \Omega)$ as a function of flux F. The function characterizes the split between detected and sub-threshold sources by the source detection network r_1, providing the probability that a source with flux F and at position Ω would be detected by the ratio estimator in Equation (6.2).

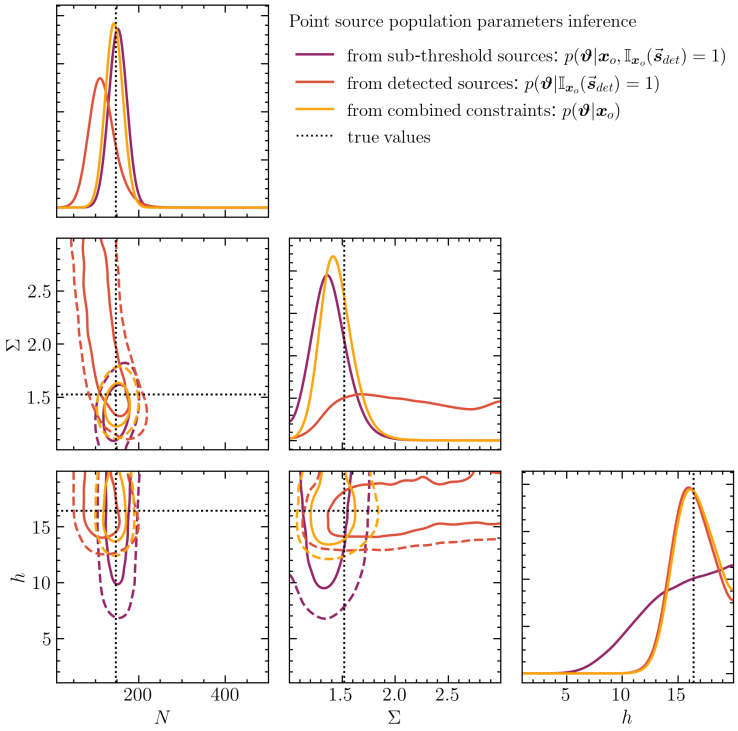

Figure 6.4: Marginal posteriors inferred by TMNRE on target observation $\boldsymbol{x}_o$ for population parameters $\boldsymbol{\theta}$. We show constraints from sub-threshold sources (violet), detected sources (orange), and combined results (yellow). In the 2D marginal posterior we indicate the 68% and 95% credible regions with solid and dashed lines respectively. For more details see Section 6.3.

sources realisations, always re-sampling the parameter Σ (see Section 6.2.1 for the hierarchical model details).

6.4 Discussion and conclusions

We have introduced a novel method to self-consistently perform point sources detection and source population parameters inference using TMNRE. The key realization underlying this methodology is that point source detection is equivalent to prior truncation and essential to reduce training data variance. With this approach, we can exploit information of detected as well as sub-threshold sources for population-level parameter inference. Detection biases are automatically accounted for in our approach. Exemplary results of our approach are presented in Figure 6.4, where we show inference results on source population parameters from both detected point sources and sub-threshold sources separately, as well as their combination.

Since the proposed method is essentially a specific implementation of TMNRE, we expect that it inherits its positive properties in terms of simulation-efficiency and scalability [160]. A possible shortcoming of this approach is that multiple neural networks need to be trained self-consistently, which on the other hand have a clear interpretation in terms of traditional analysis pipeline components. A potential application beyond those directly intended is to detectable and sub-threshold substructures in strong gravitational lenses.

7 Conclusions

Today, inference and statistics are crucial to the scientific process. Over a century ago, Ernest Rutherford is reported to have said: "If you need statistics, you did the wrong experiment" [359]. While this sentiment may have been somewhat valid in the past, it no longer aligns with the current state of the scientific process. Modern research often involves complex models designed to explain subtle phenomena within multidimensional datasets. Additionally, the volume of data available to astrophysicists and cosmologists has expanded dramatically over the past two decades and continues to increase. Consequently, our statistical analysis tools must evolve in sophistication to match this growth. Now, the primary constraint on our understanding of the universe is not the volume of data we possess, but our capacity to analyze, interpret, and derive meaningful insights from it [24].

Simulation-based inference techniques offer a qualitative shift in our approach to statistical inference and a promising avenue to fully exploit the data potential. The present work highlights several key benefits of this framework, and argues for its central role in the modern physics data analysis toolkit. The significance of the three applications presented in this thesis is underscored not only by the complexity of the statistical inference challenges they tackle, but also by the critical scientific questions they seek to resolve. Specifically, simulation-based inference can solve *hierarchical inference* of dark matter substructure population parameters from a dataset of strong lensing images, constraining dark matter nature, while fully accounting from the uncertainties coming from all the building blocks of the model (i.e., lens, source, and substructure population). It also can solve high-dimensional inference at the *field-level*, providing a means to optimally extract information from upcoming cosmological surveys. Lastly, simulation-based inference allows, for the first time to the authors' knowledge, for *self-consistent* measurement of population parameters based on detected and undetected objects in sky-maps, by making observational biases related to the point source detection part of the model itself. These applications demonstrate the versatility of simulation-based inference in tackling diverse and intricate problems, setting the stage for its broader adoption and further innovation in future research.

We are already witnessing a widespread adoption and advancement of simulation-based inference techniques for cosmological and astrophysical data analysis. However, these techniques have yet to realize their full potential, and numerous challenges remain open. Key questions include: Can the selection of optimal neural networks architectures be automatized? How to handle situations with high volume data? What is the most efficient way to sample from constrained likelihood regions in very high dimensions for sequential inference? How to reliably perform goodness-of-fit tests and identify model misspecification? Crucially, the further development of these techniques must be accompanied by the development of robust performance tests and diagnostics. All these open threads, and possibly more, are active areas of research within the community.

Our pursuit to develop powerful, fast, and robust analysis techniques is especially urgent because of the ever-increasing influx and quality of data from current and forthcoming observatories. The potential for using these observations to uncover new physics in the dark sector are therefore very bright. Now is absolutely an opportune time to explore innovative ways of maximizing the discovery potentials of these datasets. While the approaches presented in this thesis are far from the only interesting avenues for analyzing this wealth of data for new physics searches, they represent concrete steps towards fully exploiting these measurements to refine our understanding of physical laws.

Bibliography

[1] N. Anau Montel, A. Coogan, C. Correa, K. Karchev, and C. Weniger, "Estimating the warm dark matter mass from strong lensing images with truncated marginal neural ratio estimation," *Mon. Not. Roy. Astron. Soc.* **518** no. 2, (2022) 2746–2760, arXiv:2205.09126 [astro-ph.CO].

[2] N. Anau Montel, J. Alvey, and C. Weniger, "Scalable inference with autoregressive neural ratio estimation," *Mon. Not. Roy. Astron. Soc.* **530** no. 4, (2024) 4107–4124, arXiv:2308.08597 [astro-ph.IM].

[3] F. List, N. Anau Montel, and C. Weniger, "Bayesian simulation-based inference for cosmological initial conditions," in *37th Conference on Neural Information Processing Systems: Workshop on Machine Learning and the Physical Sciences*. 10, 2023. arXiv:2310.19910 [astro-ph.CO].

[4] N. Anau Montel and C. Weniger, "Detection is truncation: studying source populations with truncated marginal neural ratio estimation," in *36th Conference on Neural Information Processing Systems: Workshop on Machine Learning and the Physical Sciences*. 11, 2022. arXiv:2211.04291 [astro-ph.IM].

[5] A. Coogan, N. Anau Montel, K. Karchev, M. W. Grootes, F. Nattino, and C. Weniger, "The effect of the perturber population on subhalo measurements in strong gravitational lenses," *Mon. Not. Roy. Astron. Soc.* **527** no. 1, (2024) 66–78, arXiv:2209.09918 [astro-ph.CO].

[6] K. Karchev, N. Anau Montel, A. Coogan, and C. Weniger, "Strong-lensing source reconstruction with denoising diffusion restoration models," in *36th Conference on Neural Information Processing Systems: Workshop on Machine Learning and the Physical Sciences*. 11, 2022. arXiv:2211.04365 [astro-ph.CO].

[7] C. A. Correa, M. Schaller, S. Ploeckinger, N. Anau Montel, C. Weniger, and S. Ando, "TangoSIDM: Tantalizing models of Self-Interacting Dark Matter," *Mon. Not. Roy. Astron. Soc.* **517** no. 2, (2022) 3045–3063, arXiv:2206.11298 [astro-ph.GA].

[8] C. W. Groetsch, *Inverse problems in the mathematical sciences*. Vieweg+Teubner Verlag Wiesbaden, 1993.

[9] R. C. Aster, C. H. Thurber, and B. Borchers, *Parameter Estimation and Inverse*

Problems. Academic Press, 2005.

[10] **LSST Dark Energy Science** Collaboration, A. Abate *et al.*, "Large Synoptic Survey Telescope: Dark Energy Science Collaboration," arXiv:1211.0310 [astro-ph.CO].

[11] J. D. Simon, S. Birrer, K. Bechtol, S. Chakrabarti, F.-Y. Cyr-Racine, I. Dell'Antonio, A. Drlica-Wagner, C. Fassnacht, M. Geha, D. Gilman, Y. D. Hezaveh, D. Kim, T. S. Li, L. Strigari, and T. Treu, "Testing the Nature of Dark Matter with Extremely Large Telescopes," arXiv:1903.04742 [astro-ph.CO].

[12] J. P. Gardner *et al.*, "The James Webb Space Telescope," *Space Sci. Rev.* **123** (2006) 485, arXiv:astro-ph/0606175.

[13] **Euclid Imaging** Collaboration, A. Refregier, A. Amara, T. D. Kitching, A. Rassat, R. Scaramella, and J. Weller, "Euclid Imaging Consortium Science Book," arXiv:1001.0061 [astro-ph.IM].

[14] M. Cirelli, A. Strumia, and J. Zupan, "Dark Matter," arXiv:2406.01705 [hep-ph].

[15] A. Drlica-Wagner, C. Prescod-Weinstein, H.-B. Yu, A. Albert, M. Amin, A. Banerjee, M. Baryakhtar, K. Bechtol, S. Bird, S. Birrer, T. Bringmann, R. Caputo, S. Chakrabarti, T. Y. Chen, D. Croon, F.-Y. Cyr-Racine, W. A. Dawson, C. Dvorkin, V. Gluscevic, D. Gilman, D. Grin, R. Hložek, R. K. Leane, T. S. Li, Y.-Y. Mao, J. Meyers, S. Mishra-Sharma, J. B. Muñoz, F. Munshi, E. O. Nadler, A. Parikh, K. Perez, A. H. G. Peter, S. Profumo, K. Schutz, N. Sehgal, J. D. Simon, K. Sinha, M. Valluri, and R. H. Wechsler, "Report of the topical group on cosmic probes of dark matter for snowmass 2021," arXiv:2209.08215 [hep-ph].

[16] M. J. Mortonson, D. H. Weinberg, and M. White, "Dark Energy: A Short Review," arXiv:1401.0046 [astro-ph.CO].

[17] D. Huterer and D. L. Shafer, "Dark energy two decades after: Observables, probes, consistency tests," *Rept. Prog. Phys.* **81** no. 1, (2018) 016901, arXiv:1709.01091 [astro-ph.CO].

[18] A. Boyarsky, D. Iakubovskyi, and O. Ruchayskiy, "Next decade of sterile neutrino studies," *Phys. Dark Univ.* **1** (2012) 136–154, arXiv:1306.4954 [astro-ph.CO].

[19] M. Sajjad Athar *et al.*, "Status and perspectives of neutrino physics," *Prog. Part. Nucl. Phys.* **124** (2022) 103947, arXiv:2111.07586 [hep-ph].

[20] D. Baumann, "Inflation," in *Theoretical Advanced Study Institute in Elementary Particle Physics: Physics of the Large and the Small*, pp. 523–686. 2011. arXiv:0907.5424 [hep-th].

[21] A. Achúcarro *et al.*, "Inflation: Theory and Observations," arXiv:2203.08128 [astro-ph.CO].

[22] R. Alves Batista *et al.*, "EuCAPT White Paper: Opportunities and Challenges for Theoretical Astroparticle Physics in the Next Decade," arXiv:2110.10074 [astro-ph.HE].

[23] D. Green *et al.*, "Snowmass Theory Frontier: Astrophysics and Cosmology," arXiv:2209.06854 [hep-ph].

[24] R. Trotta, "Bayesian Methods in Cosmology," 1, 2017. arXiv:1701.01467 [astro-ph.CO].

[25] L. Verde, "Statistical methods in cosmology," in *Lectures on Cosmology: Accelerated Expansion of the Universe*, pp. 147–177. Springer, 2010. arXiv:0911.3105 [astro-ph.CO].

[26] E. D. Feigelson, R. S. De Souza, E. E. Ishida, and G. J. Babu, "Twenty-first-century statistical and computational challenges in astrophysics," *Annual Review of Statistics and Its Application* **8** no. 1, (2021) 493–517, arXiv:2005.13025 [astro-ph.IM].

[27] M. Meneghetti, *Introduction to Gravitational Lensing*. Springer Cham, 11, 2016.

[28] A. Zitrin, I. Labbé, S. Belli, R. Bouwens, R. S. Ellis, G. Roberts-Borsani, D. P. Stark, P. A. Oesch, and R. Smit, "Lyman-alpha emission from a luminous z=8.68 galaxy: Implications for galaxies as tracers of cosmic reionization," *The Astrophysical Journal* **810** no. 1, (Aug., 2015) L12, arXiv:1507.02679 [astro-ph.GA].

[29] R. P. Naidu, P. A. Oesch, P. v. Dokkum, E. J. Nelson, K. A. Suess, G. Brammer, K. E. Whitaker, G. Illingworth, R. Bouwens, S. Tacchella, J. Matthee, N. Allen, R. Bezanson, C. Conroy, I. Labbe, J. Leja, E. Leonova, D. Magee, S. H. Price, D. J. Setton, V. Strait, M. Stefanon, S. Toft, J. R. Weaver, and A. Weibel, "Two remarkably luminous galaxy candidates at $z \approx 10$–12 revealed by jwst," *The Astrophysical Journal Letters* **940** no. 1, (Nov., 2022) L14, arXiv:2207.09434 [astro-ph.GA].

[30] T. Treu, K. B. Schmidt, G. B. Brammer, B. Vulcani, X. Wang, M. Bradac, M. Dijkstra, A. Dressler, A. Fontana, R. Gavazzi, A. L. Henry, A. Hoag, K.-H. Huang, T. A. Jones, P. L. Kelly, M. A. Malkan, C. Mason, L. Pentericci, B. Poggianti, M. Stiavelli, M. Trenti, and A. v. d. Linden, "The grism lens-amplified survey from space (glass). i. survey overview and first data release," *The Astrophysical Journal* **812** no. 2, (Oct., 2015) 114, arXiv:1509.00475 [astro-ph.GA].

[31] N. Dalal and C. S. Kochanek, "Direct detection of CDM substructure," *Astrophys. J.* **572** (2002) 25–33, arXiv:astro-ph/0111456.

[32] S. Vegetti, L. V. E. Koopmans, M. W. Auger, T. Treu, and A. S. Bolton, "Inference of the cold dark matter substructure mass function at z = 0.2 using strong gravitational lenses," *Mon. Not. Roy. Astron. Soc.* **442** no. 3, (2014) 2017–2035, arXiv:1405.3666 [astro-ph.GA].

[33] D. Gilman, S. Birrer, A. Nierenberg, T. Treu, X. Du, and A. Benson, "Warm dark matter chills out: constraints on the halo mass function and the free-streaming length of dark matter with eight quadruple-image strong gravitational lenses," *Mon. Not. Roy. Astron. Soc.* **491** no. 4, (2020) 6077–6101, arXiv:1908.06983 [astro-ph.CO].

[34] J.-W. Hsueh, W. Enzi, S. Vegetti, M. Auger, C. D. Fassnacht, G. Despali, L. V. E.

Koopmans, and J. P. McKean, "SHARP – VII. New constraints on the dark matter free-streaming properties and substructure abundance from gravitationally lensed quasars," *Mon. Not. Roy. Astron. Soc.* **492** no. 2, (2020) 3047–3059, arXiv:1905.04182 [astro-ph.CO].

[35] J. Nightingale and S. Dye, "Adaptive semi-linear inversion of strong gravitational lens imaging," *Mon. Not. Roy. Astron. Soc.* **452** no. 3, (July, 2015) 2940–2959, arXiv:1412.7436 [astro-ph.IM].

[36] S. Vegetti, L. V. E. Koopmans, A. Bolton, T. Treu, and R. Gavazzi, "Detection of a Dark Substructure through Gravitational Imaging," *Mon. Not. Roy. Astron. Soc.* **408** (2010) 1969, arXiv:0910.0760 [astro-ph.CO].

[37] S. Vegetti, D. J. Lagattuta, J. P. McKean, M. W. Auger, C. D. Fassnacht, and L. V. E. Koopmans, "Gravitational detection of a low-mass dark satellite at cosmological distance," *Nature* **481** (2012) 341, arXiv:1201.3643 [astro-ph.CO].

[38] Y. D. Hezaveh *et al.*, "Detection of lensing substructure using ALMA observations of the dusty galaxy SDP.81," *Astrophys. J.* **823** no. 1, (2016) 37, arXiv:1601.01388 [astro-ph.CO].

[39] S. H. Suyu, S. Huber, R. Cañameras, M. Kromer, S. Schuldt, S. Taubenberger, A. Yıldırım, V. Bonvin, J. H. H. Chan, F. Courbin, U. Nöbauer, S. A. Sim, and D. Sluse, "Holismokes: I. highly optimised lensing investigations of supernovae, microlensing objects, and kinematics of ellipticals and spirals," *Astronomy & Astrophysics* **644** (Dec., 2020) A162, arXiv:2002.08378 [astro-ph.CO].

[40] S. Birrer, A. J. Shajib, A. Galan, M. Millon, T. Treu, A. Agnello, M. Auger, G. C.-F. Chen, L. Christensen, T. Collett, F. Courbin, C. D. Fassnacht, L. V. E. Koopmans, P. J. Marshall, J.-W. Park, C. E. Rusu, D. Sluse, C. Spiniello, S. H. Suyu, S. Wagner-Carena, K. C. Wong, M. Barnabè, A. S. Bolton, O. Czoske, X. Ding, J. A. Frieman, and L. Van de Vyvere, "Tdcosmo: Iv. hierarchical time-delay cosmography – joint inference of the hubble constant and galaxy density profiles," *Astronomy & Astrophysics* **643** (Nov., 2020) A165, arXiv:2007.02941 [astro-ph.CO].

[41] K. Karchev, A. Coogan, and C. Weniger, "Strong-lensing source reconstruction with variationally optimized Gaussian processes," *Mon. Not. Roy. Astron. Soc.* **512** no. 1, (2022) 661–685, arXiv:2105.09465 [astro-ph.IM].

[42] **EUCLID** Collaboration, R. Laureijs *et al.*, "Euclid Definition Study Report," arXiv:1110.3193 [astro-ph.CO].

[43] J. R. Bond, L. Kofman, and D. Pogosyan, "How filaments are woven into the cosmic web," *Nature* **380** (1996) 603–606, arXiv:astro-ph/9512141.

[44] J. Jasche and B. D. Wandelt, "Bayesian physical reconstruction of initial conditions from large scale structure surveys," *Mon. Not. Roy. Astron. Soc.* **432** (2013) 894, arXiv:1203.3639 [astro-ph.CO].

[45] D. A. Kniffen, C. E. Fichtel, R. C. Hartman, R. C. Lamb, and D. J. Thompson, "SAS-2 Observations of High-Energy Gamma Rays from Discrete Sources," in *Recent*

Advances in Gamma-Ray Astronomy, R. D. Wills and B. Battrick, eds., vol. 124 of *ESA Special Publication*, p. 45. July, 1977.

[46] C. E. Fichtel, R. C. Hartman, D. A. Kniffen, D. J. Thompson, G. F. Bignami, H. Ögelman, M. E. Özel, and T. Tümer, "High-energy gamma-ray results from the second Small Astronomy Satellite," *The Astrophysical Journal* **198** (May, 1975) 163–182.

[47] J.-M. Casandjian and I. A. Grenier, "A revised catalogue of EGRET gamma-ray sources," *Astron. Astrophys.* **489** (2008) 849, arXiv:0806.0113 [astro-ph].

[48] **Fermi-LAT** Collaboration, S. Abdollahi *et al.*, "Incremental Fermi Large Area Telescope Fourth Source Catalog," *Astrophys. J. Supp.* **260** no. 2, (2022) 53, arXiv:2201.11184 [astro-ph.HE].

[49] G. Dobler, D. P. Finkbeiner, I. Cholis, T. Slatyer, and N. Weiner, "The fermi haze: a gamma-ray counterpart to the microwave haze," *The Astrophysical Journal* **717** no. 2, (2010) 825, arXiv:0910.4583 [astro-ph.HE].

[50] L. Goodenough and D. Hooper, "Possible Evidence For Dark Matter Annihilation In The Inner Milky Way From The Fermi Gamma Ray Space Telescope," arXiv:0910.2998 [hep-ph].

[51] K. P. Murphy, *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.

[52] K. Cranmer, J. Brehmer, and G. Louppe, "The frontier of simulation-based inference," *Proc. Nat. Acad. Sci.* **117** no. 48, (2020) 30055–30062, arXiv:1911.01429 [stat.ML].

[53] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* **521** (05, 2015) 436–44.

[54] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind, "Automatic differentiation in machine learning: a survey," *Journal of machine learning research* **18** no. 153, (2018) 1–43, arXiv:1502.05767 [cs.SC].

[55] J. Alsing, T. Charnock, S. Feeney, and B. Wandelt, "Fast likelihood-free cosmology with neural density estimators and active learning," *Mon. Not. Roy. Astron. Soc.* **488** no. 3, (2019) 4440–4458, arXiv:1903.00007 [astro-ph.CO].

[56] A. Tejero-Cantero, J. Boelts, M. Deistler, J.-M. Lueckmann, C. Durkan, P. J. Gonçalves, D. S. Greenberg, and J. H. Macke, "sbi: A toolkit for simulation-based inference," *Journal of Open Source Software* **5** no. 52, (2020) 2505. https://doi.org/10.21105/joss.02505.

[57] B. K. Miller, A. Cole, C. Weniger, F. Nattino, O. Ku, and M. W. Grootes, "swyft: Truncated marginal neural ratio estimation in python," *Journal of Open Source Software* **7** no. 75, (2022) 4205. https://doi.org/10.21105/joss.04205.

[58] F. Rozet and A. Delaunoy, "LAMPE: Likelihood-free AMortized Posterior Estimation," 2022. https://github.com/probabilists/lampe.

[59] M. Dax, S. R. Green, J. Gair, J. H. Macke, A. Buonanno, and B. Schölkopf, "Real-Time Gravitational Wave Science with Neural Posterior Estimation," *Phys. Rev. Lett.* **127** no. 24, (2021) 241103, arXiv:2106.12594 [gr-qc].

[60] M. Crisostomi, K. Dey, E. Barausse, and R. Trotta, "Neural posterior estimation with guaranteed exact coverage: The ringdown of GW150914," *Phys. Rev. D* **108** no. 4, (2023) 044029, arXiv:2305.18528 [gr-qc].

[61] A. Kolmus, J. Janquart, T. Baka, T. van Laarhoven, C. V. D. Broeck, and T. Heskes, "Tuning neural posterior estimation for gravitational wave inference," arXiv:2403.02443 [astro-ph.IM].

[62] A. Dimitriou, D. G. Figueroa, and B. Zaldivar, "Fast Likelihood-free Reconstruction of Gravitational Wave Backgrounds," arXiv:2309.08430 [astro-ph.IM].

[63] I. M. Vílchez and C. F. Sopuerta, "Efficient Massive Black Hole Binary parameter estimation for LISA using Sequential Neural Likelihood," arXiv:2406.00565 [gr-qc].

[64] U. Bhardwaj, J. Alvey, B. K. Miller, S. Nissanke, and C. Weniger, "Sequential simulation-based inference for gravitational wave signals," *Phys. Rev. D* **108** no. 4, (2023) 042004, arXiv:2304.02035 [gr-qc].

[65] J. Alvey, U. Bhardwaj, S. Nissanke, and C. Weniger, "What to do when things get crowded? Scalable joint analysis of overlapping gravitational wave signals," arXiv:2308.06318 [gr-qc].

[66] J. Alvey, U. Bhardwaj, V. Domcke, M. Pieroni, and C. Weniger, "Simulation-based inference for stochastic gravitational wave background data analysis," *Phys. Rev. D* **109** no. 8, (2024) 083008, arXiv:2309.07954 [gr-qc].

[67] **LSST Dark Energy Science** Collaboration, S. Wagner-Carena, J. W. Park, S. Birrer, P. J. Marshall, A. Roodman, and R. H. Wechsler, "Hierarchical Inference with Bayesian Neural Networks: An Application to Strong Gravitational Lensing," *Astrophys. J.* **909** no. 2, (2021) 187, arXiv:2010.13787 [astro-ph.CO].

[68] S. Wagner-Carena, J. Aalbers, S. Birrer, E. O. Nadler, E. Darragh-Ford, P. J. Marshall, and R. H. Wechsler, "From Images to Dark Matter: End-to-end Inference of Substructure from Hundreds of Strong Gravitational Lenses," *Astrophys. J.* **942** no. 2, (2023) 75, arXiv:2203.00690 [astro-ph.CO].

[69] S. Wagner-Carena, J. Lee, J. Pennington, J. Aalbers, S. Birrer, and R. H. Wechsler, "A strong gravitational lens is worth a thousand dark matter halos: Inference on small-scale structure using sequential methods," arXiv:2404.14487 [astro-ph.CO].

[70] J. Brehmer, S. Mishra-Sharma, J. Hermans, G. Louppe, and K. Cranmer, "Mining for Dark Matter Substructure: Inferring subhalo population properties from strong lenses with machine learning," *Astrophys. J.* **886** no. 1, (2019) 49, arXiv:1909.02005 [astro-ph.CO].

[71] G. Zhang, S. Mishra-Sharma, and C. Dvorkin, "Inferring subhalo effective density

slopes from strong lensing observations with neural likelihood-ratio estimation," *Mon. Not. Roy. Astron. Soc.* **517** no. 3, (2022) 4317–4326, arXiv:2208.13796 [astro-ph.CO].

[72] B. Tucci and F. Schmidt, "EFTofLSS meets simulation-based inference: σ_8 from biased tracers," *JCAP* **05** (2024) 063, arXiv:2310.03741 [astro-ph.CO].

[73] C. Modi and O. H. E. Philcox, "Hybrid SBI or How I Learned to Stop Worrying and Learn the Likelihood," arXiv:2309.10270 [astro-ph.CO].

[74] T. L. Makinen, T. Charnock, J. Alsing, and B. D. Wandelt, "Lossless, scalable implicit likelihood inference for cosmological fields," *Journal of Cosmology and Astroparticle Physics* **11** no. 11, (2021) 049, arXiv:2107.07405 [astro-ph.CO]. [Erratum: JCAP 04, E02 (2023)].

[75] **DES** Collaboration, N. Jeffrey *et al.*, "Dark Energy Survey Year 3 results: likelihood-free, simulation-based wCDM inference with neural compression of weak-lensing map statistics," arXiv:2403.02314 [astro-ph.CO].

[76] N. Jeffrey, J. Alsing, and F. Lanusse, "Likelihood-free inference with neural compression of DES SV weak lensing map statistics," *Mon. Not. Roy. Astron. Soc.* **501** no. 1, (Nov, 2020) 954–969, arXiv:2009.08459 [stat.CO].

[77] M. von Wietersheim-Kramsta, K. Lin, N. Tessore, B. Joachimi, A. Loureiro, R. Reischke, and A. H. Wright, "KiDS-SBI: Simulation-Based Inference Analysis of KiDS-1000 Cosmic Shear," arXiv:2404.15402 [astro-ph.CO].

[78] A. Cole, B. K. Miller, S. J. Witte, M. X. Cai, M. W. Grootes, F. Nattino, and C. Weniger, "Fast and credible likelihood-free cosmology with truncated marginal neural ratio estimation," *JCAP* **09** (2022) 004, arXiv:2111.08030 [astro-ph.CO].

[79] G. Franco Abellán, G. C. n. Herrera, M. Martinelli, O. Savchenko, D. Sciotti, and C. Weniger, "Fast likelihood-free inference in the LSS Stage IV era," arXiv:2403.14750 [astro-ph.CO].

[80] A. Saxena, A. Cole, S. Gazagnes, P. D. Meerburg, C. Weniger, and S. J. Witte, "Constraining the X-ray heating and reionization using 21-cm power spectra with Marginal Neural Ratio Estimation," *Mon. Not. Roy. Astron. Soc.* **525** no. 4, (2023) 6097–6111, arXiv:2303.07339 [astro-ph.CO].

[81] K. Karchev, R. Trotta, and C. Weniger, "SICRET: Supernova Ia Cosmology with truncated marginal neural Ratio EsTimation," *Mon. Not. Roy. Astron. Soc.* **520** no. 1, (2023) 1056–1072, arXiv:2209.06733 [astro-ph.IM].

[82] K. Karchev, M. Grayling, B. M. Boyd, R. Trotta, K. S. Mandel, and C. Weniger, "SIDE-real: Supernova Ia Dust Extinction with truncated marginal neural ratio estimation applied to real data," *Mon. Not. Roy. Astron. Soc.* **530** no. 4, (2024) 3881–3896, arXiv:2403.07871 [astro-ph.CO].

[83] D. Barret and S. Dupourqué, "Simulation-Based Inference with Neural Posterior Estimation applied to X-ray spectral fitting: Demonstration of working principles

down to the Poisson regime," arXiv:2401.06061 [astro-ph.IM].

[84] M. Vasist, F. Rozet, O. Absil, P. Mollière, E. Nasedkin, and G. Louppe, "Neural posterior estimation for exoplanetary atmospheric retrieval," *Astronomy & Astrophysics* **672** (2023) A147, arXiv:2301.06575 [astro-ph.EP].

[85] C. Hahn and P. Melchior, "Accelerated Bayesian SED Modeling Using Amortized Neural Posterior Estimation," *Astrophys. J.* **938** no. 1, (2022) 11, arXiv:2203.07391 [astro-ph.GA].

[86] G. Khullar, B. Nord, A. Ćiprijanović, J. Poh, and F. Xu, "Digs: deep inference of galaxy spectra with neural posterior estimation," *Machine Learning: Science and Technology* **3** no. 4, (2022) 04LT04, arXiv:2211.09126 [astro-ph.GA].

[87] S. Mishra-Sharma and K. Cranmer, "Neural simulation-based inference approach for characterizing the Galactic Center γ-ray excess," *Phys. Rev. D* **105** no. 6, (2022) 063017, arXiv:2110.06931 [astro-ph.HE].

[88] K. Christy, E. J. Baxter, and J. Kumar, "Applying Simulation-Based Inference to Spectral and Spatial Information from the Galactic Center Gamma-Ray Excess," arXiv:2402.04549 [astro-ph.HE].

[89] J. Hermans, N. Banik, C. Weniger, G. Bertone, and G. Louppe, "Towards constraining warm dark matter with stellar streams through neural simulation-based inference," *Mon. Not. Roy. Astron. Soc.* **507** no. 2, (2021) 1999–2011, arXiv:2011.14923 [astro-ph.GA].

[90] J. Alvey, M. Gerdes, and C. Weniger, "Albatross: a scalable simulation-based inference pipeline for analysing stellar streams in the Milky Way," *Mon. Not. Roy. Astron. Soc.* **525** no. 3, (2023) 3662–3681, arXiv:2304.02032 [astro-ph.GA].

[91] J. Berteaud, C. Eckner, F. Calore, M. Clavel, and D. Haggard, "Simulation-based inference of radio millisecond pulsars in globular clusters," arXiv:2405.15691 [astro-ph.HE].

[92] S. Mishra-Sharma, "Inferring dark matter substructure with astrometric lensing beyond the power spectrum," *Mach. Learn. Sci. Tech.* **3** no. 1, (2022) 01LT03, arXiv:2110.01620 [astro-ph.CO].

[93] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *J. Chem. Phys.* **21** (1953) 1087–1092.

[94] W. K. Hastings, "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika* **57** (1970) 97–109.

[95] J. Skilling, "Nested sampling for general Bayesian computation," *Bayesian Analysis* **1** no. 4, (2006) 833–859.

[96] F. Feroz, M. P. Hobson, and M. Bridges, "MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics," *Mon. Not. Roy. Astron. Soc.* **398** (2009) 1601–1614, arXiv:0809.3437 [astro-ph].

[97] G. Ashton *et al.*, "Nested sampling for physical scientists," *Nature* **2** (2022) , arXiv:2205.15570 [stat.CO].

[98] A. F. M. Smith and G. O. Roberts, "Bayesian computation via the gibbs sampler and related markov chain monte carlo methods," *Journal of the Royal Statistical Society: Series B (Methodological)* **55** no. 1, (1993) 3–23.

[99] R. M. Neal, "Slice sampling," arXiv:physics/0009028 [physics.data-an].

[100] W. J. Handley, M. P. Hobson, and A. N. Lasenby, "PolyChord: nested sampling for cosmology," *Mon. Not. Roy. Astron. Soc.* **450** no. 1, (2015) L61–L65, arXiv:1502.01856 [astro-ph.CO].

[101] M. Betancourt, "A conceptual introduction to hamiltonian monte carlo," arXiv:1701.02434 [stat.ME].

[102] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, "Hybrid Monte Carlo," *Physics Letters B* **195** no. 2, (Sept., 1987) 216–222.

[103] R. M. Neal, "Mcmc using hamiltonian dynamics," arXiv:1206.1901 [stat.CO].

[104] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *Journal of Machine Learning Research* (2013) , arXiv:1206.7051 [stat.ML].

[105] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt, "Advances in variational inference," *IEEE transactions on pattern analysis and machine intelligence* **41** no. 8, (2018) 2008–2026, arXiv:1711.05597 [cs.LG].

[106] C. Stone, A. Adam, A. Coogan, M. J. Yantovski-Barth, A. Filipp, L. Setiawan, C. Core, R. Legin, C. Wilson, G. Missael Barco, Y. Hezaveh, and L. Perreault-Levasseur, "Caustics: A Python Package for Accelerated Strong Gravitational Lensing Simulations," arXiv:2406.15542 [astro-ph.IM].

[107] M. Morvan, A. Tsiaras, N. Nikolaou, and I. P. Waldmann, "Pylightcurve-torch: a transit modeling package for deep learning applications in pytorch," *Publications of the Astronomical Society of the Pacific* **133** no. 1021, (Mar, 2021) 034505, arXiv:2011.02030 [astro-ph.EP].

[108] J. Alvey, M. Gerdes, and C. Weniger, "sstrax: Fast stellar stream modelling in JAX." Astrophysics source code library, record ascl:2306.008, June, 2023.

[109] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Curran Associates, Inc., 2019. arXiv:1912.01703 [cs.LG].

[110] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, "JAX: composable transformations of Python+NumPy programs," 2018. http://github.com/google/jax.

[111] J. Regier, K. Fischer, K. Pamnany, A. Noack, J. Revels, M. Lam, S. Howard, R. Giordano, D. Schlegel, J. McAuliffe, *et al.*, "Cataloging the visible universe through bayesian inference in julia at petascale," *Journal of Parallel and Distributed Computing* **127** (2019) 89–104, arXiv:1801.10277 [cs.DC].

[112] R. Liu, J. D. McAuliffe, J. Regier, L. D. E. S. Collaboration, *et al.*, "Variational inference for deblending crowded starfields," *Journal of Machine Learning Research* **24** no. 179, (2023) 1–36, arXiv:2102.02409 [astro-ph.IM].

[113] S. Mishra-Sharma and K. Cranmer, "Semi-parametric γ-ray modeling with Gaussian processes and variational inference," in *34th Conference on Neural Information Processing Systems*. 10, 2020. arXiv:2010.10450 [astro-ph.HE].

[114] R. Leike, M. Glatzle, and T. Enßlin, "Resolving nearby dust clouds," *Astronomy & Astrophysics* **639** (2020) A138, arXiv:2004.06732 [astro-ph.GA].

[115] F. Leclercq, "Bayesian optimization for likelihood-free cosmological inference," *Phys. Rev. D* **98** no. 6, (2018) 063511, arXiv:1805.07152 [astro-ph.CO].

[116] A. Mootoovaloo, A. F. Heavens, A. H. Jaffe, and F. Leclercq, "Parameter Inference for Weak Lensing using Gaussian Processes and MOPED," *Mon. Not. Roy. Astron. Soc.* **497** no. 2, (2020) 2213–2226, arXiv:2005.06551 [astro-ph.CO].

[117] M. Gerbino, M. Lattanzi, M. Migliaccio, L. Pagano, L. Salvati, L. Colombo, A. Gruppuso, P. Natoli, and G. Polenta, "Likelihood methods for CMB experiments," *Front. in Phys.* **8** (2020) 15, arXiv:1909.09375 [astro-ph.CO].

[118] S. Mishra-Sharma, N. L. Rodd, and B. R. Safdi, "NPTFit: A code package for Non-Poissonian Template Fitting," *Astron. J.* **153** no. 6, (2017) 253, arXiv:1612.03173 [astro-ph.HE].

[119] S. Vegetti and L. V. E. Koopmans, "Bayesian Strong Gravitational-Lens Modelling on Adaptive Grids: Objective Detection of Mass Substructure in Galaxies," *Mon. Not. Roy. Astron. Soc.* **392** (2009) 945, arXiv:0805.0201 [astro-ph].

[120] B. K. Miller, A. Cole, G. Louppe, and C. Weniger, "Simulation-efficient marginal posterior estimation with swyft: stop wasting your precious time," in *36th Conference on Neural Information Processing Systems*. 11, 2020. arXiv:2011.13951 [astro-ph.IM].

[121] K. K. Boddy *et al.*, "Snowmass2021 theory frontier white paper: Astrophysical and cosmological probes of dark matter," *JHEAp* **35** (2022) 112–138, arXiv:2203.06380 [hep-ph].

[122] S. A. Sisson, Y. Fan, and M. A. Beaumont, "Overview of approximate bayesian computation," arXiv:1802.09720 [stat.CO].

[123] P. J. Diggle and R. J. Gratton, "Monte carlo methods of inference for implicit statistical models," *Journal of the Royal Statistical Society Series B: Statistical Methodology* **46** no. 2, (1984) 193–212.

[124] J.-M. Lueckmann, J. Boelts, D. S. Greenberg, P. J. Gonçalves, and J. H. Macke,

"Benchmarking simulation-based inference," arXiv:2101.04653 [stat.ML].

[125] D. B. Rubin, "Bayesianly justifiable and relevant frequency calculations for the applied statistician," *The Annals of Statistics* (1984) 1151–1172.

[126] J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder, "Approximate bayesian computational methods," *Statistics and computing* **22** no. 6, (2012) 1167–1180, arXiv:1101.0955 [stat.CO].

[127] C. Grazian and Y. Fan, "A review of approximate bayesian computation methods via density estimation: inference for simulator-models," arXiv:1909.02736 [stat.CO].

[128] S. M. Stigler, "Darwin, Galton and the statistical enlightenment," *Journal of the Royal Statistical Society Series A: Statistics in Society* **173** no. 3, (2010) 469–482.

[129] F. Galton, *Natural inheritance*, vol. 42. Macmillan, 1889.

[130] E. Cameron and A. Pettitt, "Approximate bayesian computation for astronomical model analysis: a case study in galaxy demographics and morphological transformation at high redshift," *Mon. Not. Roy. Astron. Soc.* **425** no. 1, (2012) 44–65, arXiv:1202.1426 [astro-ph.IM].

[131] C. Hahn, M. Vakili, K. Walsh, A. P. Hearin, D. W. Hogg, and D. Campbell, "Approximate bayesian computation in large-scale structure: constraining the galaxy–halo connection," *Mon. Not. Roy. Astron. Soc.* **469** no. 3, (2017) 2791–2805, arXiv:1607.01782 [astro-ph.CO].

[132] F. B. Davies, J. F. Hennawi, A.-C. Eilers, and Z. Lukić, "A New Method to Measure the Post-Reionization Ionizing Background from the Joint Distribution of Lyman-α and Lyman-β Forest Transmission," *Astrophys. J.* **855** no. 2, (2018) 106, arXiv:1703.10174 [astro-ph.CO].

[133] D. C. Hsu, E. B. Ford, D. Ragozzine, and R. C. Morehead, "Improving the accuracy of planet occurrence rates from kepler using approximate bayesian computation," *The Astronomical Journal* **155** no. 5, (2018) 205.

[134] D. A. Riechers *et al.*, "COLDz: Shape of the CO Luminosity Function at High Redshift and the Cold Gas History of the Universe," *Astrophys. J.* **872** no. 1, (2019) 7, arXiv:1808.04371 [astro-ph.GA].

[135] J. Cisewski-Kehe, G. Weller, and C. Schafer, "A preferential attachment model for the stellar initial mass function," *Electronic Journal of Statistics* **13** no. 1, (2019) 1580–1607, arXiv:1904.11306 [astro-ph.IM].

[136] J. Akeret, A. Refregier, A. Amara, S. Seehars, and C. Hasner, "Approximate Bayesian Computation for Forward Modeling in Cosmology," *JCAP* **08** (2015) 043, arXiv:1504.07245 [astro-ph.CO].

[137] E. Jennings and M. Madigan, "astroabc: an approximate bayesian computation sequential monte carlo sampler for cosmological parameter estimation," *Astronomy and computing* **19** (2017) 16–22, arXiv:1608.07606 [astro-ph.IM].

[138] **COIN** Collaboration, E. E. O. Ishida, S. D. P. Vitenti, M. Penna-Lima, J. Cisewski, R. S. de Souza, A. M. M. Trindade, E. Cameron, and V. C. Busti, "*COSMOABC*: Likelihood-free inference via Population Monte Carlo Approximate Bayesian Computation," *Astron. Comput.* **13** (2015) 1–11, arXiv:1504.06129 [astro-ph.CO].

[139] A. Weyant, C. Schafer, and W. M. Wood-Vasey, "Likelihood-Free Cosmological Inference with Type Ia Supernovae: Approximate Bayesian Computation for a Complete Treatment of Uncertainty," *Astrophys. J.* **764** (2013) 116, arXiv:1206.2563 [astro-ph.CO].

[140] E. J. Baxter, J. G. Christy, and J. Kumar, "Approximate Bayesian Computation applied to the Diffuse Gamma-Ray Sky," *Mon. Not. Roy. Astron. Soc.* **516** no. 2, (2022) 2326–2336, arXiv:2112.00255 [astro-ph.HE].

[141] N. Bourriche and F. Capel, "Beyond the local void: A comprehensive view on the origins of the amaterasu particle," arXiv:2406.16483 [astro-ph.HE].

[142] J. Brehmer, K. Cranmer, G. Louppe, and J. Pavez, "A guide to constraining effective field theories with machine learning," *Physical Review D* **98** no. 5, (2018) 052004, arXiv:1805.00020 [hep-ph].

[143] J. Brehmer and K. Cranmer, "Simulation-based inference methods for particle physics," arXiv:2010.06439 [hep-ph].

[144] G. Papamakarios, "Neural density estimation and likelihood-free inference," arXiv:1910.13233 [stat.ML].

[145] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," *Journal of Machine Learning Research* **22** no. 57, (2021) 1–64, arXiv:1912.02762 [stat.ML].

[146] G. Papamakarios and I. Murray, "Fast ϵ-free Inference of Simulation Models with Bayesian Conditional Density Estimation," arXiv:1605.06376 [stat.ML].

[147] M. Ho *et al.*, "LtU-ILI: An All-in-One Framework for Implicit Inference in Astrophysics and Cosmology," *Open J. Astrophys.* **7** (2024) 001c.120559, arXiv:2402.05137 [astro-ph.IM].

[148] D. S. Greenberg, M. Nonnenmacher, and J. H. Macke, "Automatic posterior transformation for likelihood-free inference," arXiv:1905.07488 [cs.LG].

[149] G. Papamakarios, D. C. Sterratt, and I. Murray, "Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows," arXiv:1805.07226 [stat.ML].

[150] C. Durkan, G. Papamakarios, and I. Murray, "Sequential neural methods for likelihood-free inference," arXiv:1811.08723 [stat.ML].

[151] K. Cranmer, J. Pavez, and G. Louppe, "Approximating likelihood ratios with calibrated discriminative classifiers," arXiv:1506.02169 [stat.AP].

[152] M. U. Gutmann, R. Dutta, S. Kaski, and J. Corander, "Likelihood-free inference via

classification," *Statistics and Computing* **28** (2018) 411–425, arXiv:1407.4981 [stat.CO].

[153] J. Hermans, V. Begy, and G. Louppe, "Likelihood-free MCMC with Amortized Approximate Ratio Estimators," arXiv:1903.04057 [stat.ML].

[154] O. Thomas, R. Dutta, J. Corander, S. Kaski, and M. U. Gutmann, "Likelihood-free inference by ratio estimation," *Bayesian Analysis* **17** no. 1, (2022) 1–31, arXiv:1611.10242 [stat.ML].

[155] S. Kullback and R. A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics* **22** no. 1, (1951) 79–86.

[156] I. Kobyzev, S. J. Prince, and M. A. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE transactions on pattern analysis and machine intelligence* **43** no. 11, (2020) 3964–3979, arXiv:1908.09257 [stat.ML].

[157] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," in *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis, eds., vol. 20. Curran Associates, Inc., 2007.

[158] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980 [cs.LG].

[159] L. Ambrogioni, U. Güçlü, J. Berezutskaya, E. Borne, Y. Güçlütürk, M. Hinne, E. Maris, and M. Gerven, "Forward amortized inference for likelihood-free variational marginalization," in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 777–786, PMLR. 2019. arXiv:1805.11542 [stat.ML].

[160] B. K. Miller, A. Cole, P. Forré, G. Louppe, and C. Weniger, "Truncated marginal neural ratio estimation," in *37th Conference on Neural Information Processing Systems*. 07, 2021. arXiv:2107.01214 [stat.ML].

[161] B. K. Miller, A. Cole, C. Weniger, F. Nattino, O. Ku, and M. W. Grootes, "swyft: Truncated Marginal Neural Ratio Estimation in Python," *J. Open Source Softw.* **7** no. 75, (2022) 4205.

[162] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, vol. 31 of *Stochastic Modelling and Applied Probability*. Springer, 1996.

[163] A. Mao, M. Mohri, and Y. Zhong, "Cross-entropy loss functions: Theoretical analysis and applications," in *International conference on Machine learning*, pp. 23803–23828, PMLR. 2023. arXiv:2304.07288 [cs.LG].

[164] D. Piras, A. Polanska, A. Spurio Mancini, M. A. Price, and J. D. McEwen, "The future of cosmological likelihood-based inference: accelerated high-dimensional parameter estimation and model comparison," arXiv:2405.12965 [astro-ph.CO].

[165] N. Jeffrey and B. D. Wandelt, "Solving high-dimensional parameter inference: marginal posterior densities & Moment Networks," in *34th Conference on Neural Information Processing Systems*. 11, 2020. arXiv:2011.05991 [stat.ML].

[166] J.-M. Lueckmann, P. J. Goncalves, G. Bassetto, K. Öcal, M. Nonnenmacher, and J. H. Macke, "Flexible statistical inference for mechanistic models of neural dynamics," in *31st Conference on Neural Information Processing Systems*. 11, 2017. arXiv:1711.01861 [stat.ML].

[167] C. Durkan, I. Murray, and G. Papamakarios, "On contrastive learning for likelihood-free inference," in *37th Conference on Neural Information Processing Systems*. 02, 2020. arXiv:2002.03712 [stat.ML].

[168] M. Deistler, P. J. Goncalves, and J. H. Macke, "Truncated proposals for scalable and hassle-free simulation-based inference," in *36th Conference on Neural Information Processing Systems*. 10, 2022. arXiv:2210.04815 [stat.ML].

[169] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778. 2016.

[170] V. Roy, "Convergence diagnostics for markov chain monte carlo," *Annual Review of Statistics and Its Application* **7** no. Volume 7, 2020, (2020) 387–412, arXiv:1909.11827 [stat.CO].

[171] J. Hermans, A. Delaunoy, F. Rozet, A. Wehenkel, V. Begy, and G. Louppe, "A Trust Crisis In Simulation-Based Inference? Your Posterior Approximations Can Be Unfaithful," *Transactions on Machine Learning Research* (10, 2022) , arXiv:2110.06581 [stat.ML].

[172] N. Dalmasso, R. Izbicki, and A. Lee, "Confidence sets and hypothesis testing in a likelihood-free inference setting," in *International Conference on Machine Learning*, pp. 2323–2334, PMLR. 2020. arXiv:2002.10399 [stat.ME].

[173] N. Dalmasso, L. Masserano, D. Zhao, R. Izbicki, and A. B. Lee, "Likelihood-free frequentist inference: Bridging classical statistics and machine learning for reliable simulator-based inference," arXiv:2107.03920 [stat.ML].

[174] B. Amos *et al.*, "Tutorial on amortized optimization," *Foundations and Trends® in Machine Learning* **16** no. 5, (2023) 592–732, arXiv:2202.00665 [cs.LG].

[175] J. Linhart, A. Gramfort, and P. Rodrigues, "L-c2st: Local diagnostics for posterior approximations in simulation-based inference," in *36th Conference on Neural Information Processing Systems*, vol. 36. 2024. arXiv:2306.03580 [stat.ML].

[176] P. Lemos, A. Coogan, Y. Hezaveh, and L. Perreault-Levasseur, "Sampling-based accuracy testing of posterior estimators for general inference," in *International Conference on Machine Learning*, pp. 19256–19273, PMLR. 2023. arXiv:2302.03026 [stat.ML].

[177] T. Charnock, G. Lavaux, and B. D. Wandelt, "Automatic physical inference with information maximizing neural networks," *Physical Review D* **97** no. 8, (2018) 083004, arXiv:1802.03537 [astro-ph.IM].

[178] J. Alsing and B. Wandelt, "Nuisance hardened data compression for fast

likelihood-free inference," *Mon. Not. Roy. Astron. Soc.* **488** no. 4, (2019) 5093–5103, arXiv:1903.01473 [astro-ph.CO].

[179] V. C. Rubin, N. Thonnard, and W. K. Ford, Jr., "Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 /R = 4kpc/ to UGC 2885 /R = 122 kpc/," *Astrophys. J.* **238** (1980) 471.

[180] F. Zwicky, "Die Rotverschiebung von extragalaktischen Nebeln," *Helv. Phys. Acta* **6** (1933) 110–127.

[181] **Planck** Collaboration, P. A. R. Ade *et al.*, "Planck 2015 results. XIII. Cosmological parameters," *Astron. Astrophys.* **594** (2016) A13, arXiv:1502.01589 [astro-ph.CO].

[182] A. N. Taylor, S. Dye, T. J. Broadhurst, N. Benitez, and E. van Kampen, "Gravitational lens magnification and the mass of Abell 1689," *Astrophys. J.* **501** (1998) 539, arXiv:astro-ph/9801158.

[183] P. J. E. Peebles, "Large scale background temperature and mass fluctuations due to scale invariant primeval perturbations," *Astrophys. J. Lett.* **263** (1982) L1–L5.

[184] B. Moore, S. Ghigna, F. Governato, G. Lake, T. R. Quinn, J. Stadel, and P. Tozzi, "Dark matter substructure within galactic halos," *Astrophys. J. Lett.* **524** (1999) L19–L22, arXiv:astro-ph/9907411.

[185] A. A. Klypin, A. V. Kravtsov, O. Valenzuela, and F. Prada, "Where are the missing Galactic satellites?," *Astrophys. J.* **522** (1999) 82–92, arXiv:astro-ph/9901240.

[186] W. J. G. de Blok and S. S. McGaugh, "The Dark and visible matter content of low surface brightness disk galaxies," *Mon. Not. Roy. Astron. Soc.* **290** (1997) 533–552, arXiv:astro-ph/9704274.

[187] J. S. Bullock and M. Boylan-Kolchin, "Small-Scale Challenges to the ΛCDM Paradigm," *Ann. Rev. Astron. Astrophys.* **55** (2017) 343–387, arXiv:1707.04256 [astro-ph.CO].

[188] J. S. Bullock, "Notes on the Missing Satellites Problem," arXiv:1009.4505 [astro-ph.CO].

[189] P. Colin, V. Avila-Reese, and O. Valenzuela, "Substructure and halo density profiles in a warm dark matter cosmology," *Astrophys. J.* **542** (2000) 622–630, arXiv:astro-ph/0004115.

[190] C. J. Hogan and J. J. Dalcanton, "New dark matter physics: clues from halo structure," *Phys. Rev. D* **62** (2000) 063511, arXiv:astro-ph/0002330.

[191] M. R. Lovell, C. S. Frenk, V. R. Eke, A. Jenkins, L. Gao, and T. Theuns, "The properties of warm dark matter haloes," *Mon. Not. Roy. Astron. Soc.* **439** (2014) 300–317, arXiv:1308.1399 [astro-ph.CO].

[192] D. N. Spergel and P. J. Steinhardt, "Observational evidence for selfinteracting cold dark matter," *Phys. Rev. Lett.* **84** (2000) 3760–3763, arXiv:astro-ph/9909386.

[193] W. Hu, R. Barkana, and A. Gruzinov, "Cold and fuzzy dark matter," *Phys. Rev. Lett.* **85** (2000) 1158–1161, arXiv:astro-ph/0003365.

[194] M. R. Buckley and A. H. G. Peter, "Gravitational probes of dark matter physics," *Phys. Rept.* **761** (2018) 1–60, arXiv:1712.06615 [astro-ph.CO].

[195] G. Efstathiou, "Suppressing the formation of dwarf galaxies via photoionization," *Mon. Not. Roy. Astron. Soc.* **256** (1992) 43P–47P.

[196] A. Fitts *et al.*, "FIRE in the Field: Simulating the Threshold of Galaxy Formation," *Mon. Not. Roy. Astron. Soc.* **471** no. 3, (2017) 3547–3562, arXiv:1611.02281 [astro-ph.GA].

[197] S. Adhikari *et al.*, "Astrophysical Tests of Dark Matter Self-Interactions," arXiv:2207.10638 [astro-ph.CO].

[198] C. S. Kochanek, "The saas fee lectures on strong gravitational lensing," arXiv:astro-ph/0407232.

[199] S. Vegetti *et al.*, "Strong gravitational lensing as a probe of dark matter," arXiv:2306.11781 [astro-ph.CO].

[200] **LSST Dark Matter Group** Collaboration, A. Drlica-Wagner, Y.-Y. Mao, S. Adhikari, R. Armstrong, A. Banerjee, N. Banik, K. Bechtol, S. Bird, K. K. Boddy, A. Bonaca, J. Bovy, M. R. Buckley, E. Bulbul, C. Chang, G. Chapline, J. Cohen-Tanugi, A. Cuoco, F.-Y. Cyr-Racine, W. A. Dawson, A. D. Rivero, C. Dvorkin, D. Erkal, C. D. Fassnacht, J. García-Bellido, and M. Giannotti, "Probing the fundamental nature of dark matter with the large synoptic survey telescope," arXiv:1902.01055 [astro-ph.CO].

[201] S.-d. Mao and P. Schneider, "Evidence for substructure in lens galaxies?," *Mon. Not. Roy. Astron. Soc.* **295** (1998) 587–594, arXiv:astro-ph/9707187.

[202] A. M. Nierenberg, T. Treu, S. A. Wright, C. D. Fassnacht, and M. W. Auger, "Detection of substructure with adaptive optics integral field spectroscopy of the gravitational lens B1422+231," *Mon. Not. Roy. Astron. Soc.* **442** no. 3, (2014) 2434–2445, arXiv:1402.1496 [astro-ph.GA].

[203] A. M. Nierenberg, T. Treu, G. Brammer, A. H. G. Peter, C. D. Fassnacht, C. R. Keeton, C. S. Kochanek, K. B. Schmidt, D. Sluse, and S. A. Wright, "Probing dark matter substructure in the gravitational lens HE 0435−1223 with the WFC3 grism," *Mon. Not. Roy. Astron. Soc.* **471** no. 2, (2017) 2224–2236, arXiv:1701.05188 [astro-ph.CO].

[204] D. Gilman, S. Birrer, T. Treu, C. R. Keeton, and A. Nierenberg, "Probing the nature of dark matter by forward modelling flux ratios in strong gravitational lenses," *Mon. Not. Roy. Astron. Soc.* **481** no. 1, (2018) 819–834, arXiv:1712.04945 [astro-ph.CO].

[205] J. W. Hsueh, C. D. Fassnacht, S. Vegetti, J. P. McKean, C. Spingola, M. W. Auger, L. V. E. Koopmans, and D. J. Lagattuta, "SHARP – II. Mass structure in strong

lenses is not necessarily dark matter substructure: a flux ratio anomaly from an edge-on disc in B1555+375," *Mon. Not. Roy. Astron. Soc.* **463** no. 1, (2016) L51–L55, arXiv:1601.01671 [astro-ph.CO].

[206] J. W. Hsueh, L. Oldham, C. Spingola, S. Vegetti, C. D. Fassnacht, M. W. Auger, L. V. E. Koopmans, J. P. McKean, and D. J. Lagattuta, "SHARP – IV. An apparent flux-ratio anomaly resolved by the edge-on disc in B0712+472," *Mon. Not. Roy. Astron. Soc.* **469** no. 3, (2017) 3713–3721, arXiv:1701.06575 [astro-ph.GA].

[207] L. V. E. Koopmans, "Gravitational lensing & stellar dynamics," *EAS Publ. Ser.* **20** (2006) 161, arXiv:astro-ph/0511121.

[208] S. Vegetti and L. V. E. Koopmans, "Statistics of CDM substructure from strong gravitational lensing: quantifying the mass fraction and mass function," *Mon. Not. Roy. Astron. Soc.* **400** (2009) 1583–1592, arXiv:0903.4752 [astro-ph.CO].

[209] L. V. E. Koopmans, "Gravitational-mass imaging of CDM substructure," *Mon. Not. Roy. Astron. Soc.* **363** (2005) 1136, arXiv:astro-ph/0501324.

[210] S. Vegetti, O. Czoske, and L. V. E. Koopmans., "Quantifying dwarf satellites through gravitational imaging: the case of SDSS J120602.09+514229.5," *Mon. Not. Roy. Astron. Soc.* **407** (2010) 225, arXiv:1002.4708 [astro-ph.CO].

[211] J. M. Diego, M. Pascale, B. J. Kavanagh, P. Kelly, L. Dai, B. Frye, and T. Broadhurst, "Godzilla, a monster lurks in the Sunburst galaxy," *Astron. Astrophys.* **665** (2022) A134, arXiv:2203.08158 [astro-ph.GA].

[212] S. Vegetti, G. Despali, M. R. Lovell, and W. Enzi, "Constraining sterile neutrino cosmologies with strong gravitational lensing observations at redshift z $\sim$ 0.2," *Mon. Not. Roy. Astron. Soc.* **481** no. 3, (2018) 3661–3669, arXiv:1801.01505 [astro-ph.CO].

[213] E. Ritondale, S. Vegetti, G. Despali, M. W. Auger, L. V. E. Koopmans, and J. P. McKean, "Low-mass halo perturbations in strong gravitational lenses at redshift z $\sim$ 0.5 are consistent with CDM," *Mon. Not. Roy. Astron. Soc.* **485** no. 2, (2019) 2179–2193, arXiv:1811.03627 [astro-ph.CO].

[214] J. Skilling, "Nested Sampling," *AIP Conf. Proc.* **735** no. 1, (2004) 395.

[215] J. W. Nightingale *et al.*, "Scanning for dark matter subhaloes in Hubble Space Telescope imaging of 54 strong lenses," *Mon. Not. Roy. Astron. Soc.* **527** no. 4, (2023) 10480–10506, arXiv:2209.10566 [astro-ph.CO].

[216] B. J. Brewer, D. Huijser, and G. F. Lewis, "Trans-Dimensional Bayesian Inference for Gravitational Lens Substructures," *Mon. Not. Roy. Astron. Soc.* **455** no. 2, (2016) 1819–1829, arXiv:1508.00662 [astro-ph.IM].

[217] T. Daylan, F.-Y. Cyr-Racine, A. Diaz Rivero, C. Dvorkin, and D. P. Finkbeiner, "Probing the small-scale structure in strongly lensed systems via transdimensional inference," *Astrophys. J.* **854** no. 2, (2018) 141, arXiv:1706.06111 [astro-ph.CO].

[218] A. Galan, G. Vernardos, A. Peel, F. Courbin, and J.-L. Starck, "Using wavelets to

capture deviations from smoothness in galaxy-scale strong lenses," *Astron. Astrophys.* **668** (2022) A155, arXiv:2207.05763 [astro-ph.IM].

[219] A. Diaz Rivero, F.-Y. Cyr-Racine, and C. Dvorkin, "Power spectrum of dark matter substructure in strong gravitational lenses," *Phys. Rev. D* **97** no. 2, (2018) 023001, arXiv:1707.04590 [astro-ph.CO].

[220] A. D. Rivero, C. Dvorkin, F.-Y. Cyr-Racine, J. Zavala, and M. Vogelsberger, "Gravitational lensing and the power spectrum of dark matter substructure: Insights from the ethos n-body simulations," *Phys. Rev. D* **98** (2018) 103517, arXiv:1809.00004 [astro-ph.CO].

[221] S. Brennan, A. J. Benson, F.-Y. Cyr-Racine, C. R. Keeton, L. A. Moustakas, and A. R. Pullen, "Quantifying the power spectrum of small-scale structure in semi-analytic galaxies," *Mon. Not. Roy. Astron. Soc.* **488** no. 4, (2019) 5085–5092, arXiv:1808.03501 [astro-ph.GA].

[222] A. Çağan Şengül, A. Tsang, A. Diaz Rivero, C. Dvorkin, H.-M. Zhu, and U. Seljak, "Quantifying the line-of-sight halo contribution to the dark matter convergence power spectrum from strong gravitational lenses," *Phys. Rev. D* **102** no. 6, (2020) 063502, arXiv:2006.07383 [astro-ph.CO].

[223] S. Chatterjee and L. V. E. Koopmans, "The inner mass power spectrum of galaxies using strong gravitational lensing: beyond linear approximation," *Mon. Not. Roy. Astron. Soc.* **474** no. 2, (2018) 1762–1772, arXiv:1710.03075 [astro-ph.CO].

[224] F.-Y. Cyr-Racine, C. R. Keeton, and L. A. Moustakas, "Beyond subhalos: Probing the collective effect of the universe's small-scale structure with gravitational lensing," *Phys. Rev. D* **100** (2019) 023013, arXiv:1806.07897 [astro-ph.CO].

[225] D. Bayer, S. Chatterjee, L. V. E. Koopmans, S. Vegetti, J. P. McKean, T. Treu, C. D. Fassnacht, and K. Glazebrook, "Probing sub-galactic mass structure with the power spectrum of surface-brightness anomalies in high-resolution observations of galaxy–galaxy strong gravitational lenses. II. Observational constraints on the subgalactic matter power spectrum," *Mon. Not. Roy. Astron. Soc.* **523** no. 1, (2023) 1310–1325, arXiv:1803.05952 [astro-ph.GA].

[226] S. Birrer, A. Amara, and A. Refregier, "Lensing substructure quantification in RXJ1131-1231: A 2 keV lower bound on dark matter thermal relic mass," *JCAP* **05** (2017) 037, arXiv:1702.00009 [astro-ph.CO].

[227] Q. He, A. Robertson, J. Nightingale, S. Cole, C. S. Frenk, R. Massey, A. Amvrosiadis, R. Li, X. Cao, and A. Etherington, "A forward-modelling method to infer the dark matter particle mass from strong gravitational lenses," *Mon. Not. Roy. Astron. Soc.* **511** no. 2, (2022) 3046–3062, arXiv:2010.13221 [astro-ph.CO].

[228] Y. D. Hezaveh, L. Perreault Levasseur, and P. J. Marshall, "Fast Automated Analysis of Strong Gravitational Lenses with Convolutional Neural Networks," *Nature* **548** (2017) 555–557, arXiv:1708.08842 [astro-ph.IM].

[229] L. Perreault Levasseur, Y. D. Hezaveh, and R. H. Wechsler, "Uncertainties in

Parameters Estimated with Neural Networks: Application to Strong Gravitational Lensing," *Astrophys. J. Lett.* **850** no. 1, (2017) L7, arXiv:1708.08843 [astro-ph.CO].

[230] W. R. Morningstar, L. Perreault Levasseur, Y. D. Hezaveh, R. Blandford, P. Marshall, P. Putzky, T. D. Rueter, R. Wechsler, and M. Welling, "Data-Driven Reconstruction of Gravitationally Lensed Galaxies using Recurrent Inference Machines," *The Astrophysical Journal* **883** no. 1, (Sep, 2019) 14, arXiv:1901.01359 [astro-ph.IM].

[231] G. Vernardos, G. Tsagkatakis, and Y. Pantazis, "Quantifying the structure of strong gravitational lens potentials with uncertainty-aware deep neural networks," *Mon. Not. Roy. Astron. Soc.* **499** no. 4, (Oct., 2020) 5641–5652, arXiv:2010.07314 [astro-ph.GA].

[232] A. D. Rivero and C. Dvorkin, "Direct detection of dark matter substructure in strong lens images with convolutional neural networks," *Phys. Rev. D* **101** (2020) 023515, arXiv:1910.00015 [astro-ph.CO].

[233] S. Alexander, S. Gleyzer, E. McDonough, M. W. Toomey, and E. Usai, "Deep Learning the Morphology of Dark Matter Substructure," *Astrophys. J.* **893** (2020) 15, arXiv:1909.07346 [astro-ph.CO].

[234] B. Ostdiek, A. Diaz Rivero, and C. Dvorkin, "Image segmentation for analyzing galaxy-galaxy strong lensing systems," *Astron. Astrophys.* **657** (2022) L14, arXiv:2009.06663 [astro-ph.CO].

[235] B. Ostdiek, A. Diaz Rivero, and C. Dvorkin, "Extracting the Subhalo Mass Function from Strong Lens Images with Image Segmentation," *Astrophys. J.* **927** no. 1, (2022) 83, arXiv:2009.06639 [astro-ph.CO].

[236] A. Boyarsky, M. Drewes, T. Lasserre, S. Mertens, and O. Ruchayskiy, "Sterile neutrino Dark Matter," *Prog. Part. Nucl. Phys.* **104** (2019) 1–45, arXiv:1807.07938 [hep-ph].

[237] J. R. Bond, A. S. Szalay, and M. S. Turner, "Formation of Galaxies in a Gravitino Dominated Universe," *Phys. Rev. Lett.* **48** (1982) 1636.

[238] A. S. Bolton, S. Burles, L. V. E. Koopmans, T. Treu, and L. A. Moustakas, "The sloan lens acs survey. 1. a large spectroscopically selected sample of massive early-type lens galaxies," *Astrophys. J.* **638** (2006) 703–724, arXiv:astro-ph/0511453.

[239] **BOSS** Collaboration, J. R. Brownstein *et al.*, "The BOSS Emission-Line Lens Survey (BELLS). I. A large spectroscopically selected sample of Lens Galaxies at redshift ˜ 0.5," *Astrophys. J.* **744** (2012) 41–62, arXiv:1112.3683 [astro-ph.CO].

[240] L. V. E. Koopmans, I. W. A. Browne, and N. J. Jackson, "Strong gravitational lensing with SKA," *New Astron. Rev.* **48** (2004) 1085–1094, arXiv:astro-ph/0409095.

[241] **LSST Dark Energy Science** Collaboration, B. Abolfathi *et al.*, "The LSST DESC DC2 Simulated Sky Survey," *Astrophys. J. Suppl.* **253** no. 1, (2021) 31, arXiv:2010.05926 [astro-ph.IM].

[242] T. E. Collett, "The population of galaxy-galaxy strong lenses in forthcoming optical imaging surveys," *Astrophys. J.* **811** no. 1, (2015) 20, arXiv:1507.02657 [astro-ph.CO].

[243] J. P. McKean, N. Jackson, S. Vegetti, M. Rybak, S. Serjeant, L. V. E. Koopmans, R. B. Metcalf, C. D. Fassnacht, P. J. Marshall, and M. Pandey-Pommier, "Strong gravitational lensing with the ska," arXiv:1502.03362 [astro-ph.GA].

[244] A. Coogan, K. Karchev, and C. Weniger, "Targeted Likelihood-Free Inference of Dark Matter Substructure in Strongly-Lensed Galaxies," in *34th Conference on Neural Information Processing Systems*. 10, 2020. arXiv:2010.07032 [astro-ph.CO].

[245] **Astropy** Collaboration, T. P. Robitaille *et al.*, "Astropy: A Community Python Package for Astronomy," *Astron. Astrophys.* **558** (2013) A33, arXiv:1307.6212 [astro-ph.IM].

[246] **Astropy** Collaboration, A. M. Price-Whelan *et al.*, "The Astropy Project: Building an Open-science Project and Status of the v2.0 Core Package," *Astron. J.* **156** no. 3, (2018) 123, arXiv:1801.02634 [astro-ph.IM].

[247] **Planck** Collaboration, N. Aghanim *et al.*, "Planck 2018 results. VI. Cosmological parameters," *Astron. Astrophys.* **641** (2020) A6, arXiv:1807.06209 [astro-ph.CO]. [Erratum: Astron.Astrophys. 652, C4 (2021)].

[248] M. Bartelmann and P. Schneider, "Weak gravitational lensing," *Phys. Rept.* **340** (2001) 291–472, arXiv:astro-ph/9912508.

[249] C. Giocoli, G. Tormen, R. K. Sheth, and F. C. van den Bosch, "The Substructure Hierarchy in Dark Matter Haloes," *Mon. Not. Roy. Astron. Soc.* **404** (2010) 502–517, arXiv:0911.0436 [astro-ph.CO].

[250] J. L. Tinker, A. V. Kravtsov, A. Klypin, K. Abazajian, M. S. Warren, G. Yepes, S. Gottlober, and D. E. Holz, "Toward a halo mass function for precision cosmology: The Limits of universality," *Astrophys. J.* **688** (2008) 709–728, arXiv:0803.2706 [astro-ph].

[251] J. L. Sérsic, "Influence of the atmospheric and instrumental dispersion on the brightness distribution in a galaxy," *Boletin de la Asociacion Argentina de Astronomia La Plata Argentina* **6** (Feb., 1963) 41–43.

[252] L. Ciotti and G. Bertin, "Analytical properties of the r^(1/m) luminosity law," *Astron. Astrophys.* **352** (1999) 447, arXiv:astro-ph/9911078.

[253] S. H. Suyu, P. J. Marshall, R. D. Blandford, C. D. Fassnacht, L. V. E. Koopmans, J. P. McKean, and T. Treu, "Dissecting the Gravitational Lens B1608+656: Lens Potential Reconstruction," *Astrophys. J.* **691** (2009) 277–298, arXiv:0804.2827

[astro-ph].

[254] N. Tessore and R. B. Metcalf, "The elliptical power law profile lens," *Astron. Astrophys.* **580** (2015) A79, arXiv:1507.01819 [astro-ph.CO].

[255] C. M. O'Riordan, S. J. Warren, and D. J. Mortlock, "Galaxy mass profiles from strong lensing ii: The elliptical power-law model," *Mon. Not. Roy. Astron. Soc.* **496** (04, 2020) 3424–3435, arXiv:2004.13435 [astro-ph.GA].

[256] M. Chianese, A. Coogan, P. Hofma, S. Otten, and C. Weniger, "Differentiable Strong Lensing: Uniting Gravity and Neural Nets through Differentiable Probabilistic Programming," *Mon. Not. Roy. Astron. Soc.* **496** no. 1, (2020) 381–393, arXiv:1910.06157 [astro-ph.CO].

[257] P. Schneider and D. Sluse, "Mass-sheet degeneracy, power-law models and external convergence: Impact on the determination of the Hubble constant from gravitational lensing," *Astron. Astrophys.* **559** (2013) A37, arXiv:1306.0901 [astro-ph.CO].

[258] P. Schneider and D. Sluse, "Source-position transformation – an approximate invariance in strong gravitational lensing," *Astron. Astrophys.* **564** (2014) A103, arXiv:1306.4675 [astro-ph.CO].

[259] G. Despali, S. Vegetti, S. D. M. White, C. Giocoli, and F. C. van den Bosch, "Modelling the line-of-sight contribution in substructure lensing," *Mon. Not. Roy. Astron. Soc.* **475** no. 4, (2018) 5424–5442, arXiv:1710.05029 [astro-ph.CO].

[260] Q. He *et al.*, "Galaxy–galaxy strong lens perturbations: line-of-sight haloes versus lens subhaloes," *Mon. Not. Roy. Astron. Soc.* **512** no. 4, (2022) 5862–5873, arXiv:2110.04512 [astro-ph.CO].

[261] E. A. Baltz, P. Marshall, and M. Oguri, "Analytic models of plausible gravitational lens potentials," *JCAP* **01** (2009) 015, arXiv:0705.0682 [astro-ph].

[262] J. F. Navarro, C. S. Frenk, and S. D. M. White, "A Universal density profile from hierarchical clustering," *Astrophys. J.* **490** (1997) 493–508, arXiv:astro-ph/9611107.

[263] J. Richings, C. Frenk, A. Jenkins, A. Robertson, and M. Schaller, "A high-resolution cosmological simulation of a strong gravitational lens," *Mon. Not. Roy. Astron. Soc.* **501** no. 3, (2021) 4657–4668, arXiv:2005.14495 [astro-ph.CO].

[264] N. C. Amorisco *et al.*, "Halo concentration strengthens dark matter constraints in galaxy–galaxy strong lensing analyses," *Mon. Not. Roy. Astron. Soc.* **510** no. 2, (2022) 2464–2479, arXiv:2109.00018 [astro-ph.CO].

[265] G. Despali and S. Vegetti, "The impact of baryonic physics on the subhalo mass function and implications for gravitational lensing," *Mon. Not. Roy. Astron. Soc.* **469** no. 2, (2017) 1997–2010, arXiv:1608.06938 [astro-ph.GA].

[266] V. Springel, J. Wang, M. Vogelsberger, A. Ludlow, A. Jenkins, A. Helmi, J. F. Navarro, C. S. Frenk, and S. D. M. White, "The Aquarius Project: the subhalos of galactic halos," *Mon. Not. Roy. Astron. Soc.* **391** (2008) 1685–1711,

arXiv:0809.0898 [astro-ph].

[267] P. Fleury, J. Larena, and J.-P. Uzan, "Line-of-sight effects in strong gravitational lensing," *JCAP* **08** (2021) 024, arXiv:2104.08883 [astro-ph.CO].

[268] A. Schneider, R. E. Smith, A. V. Maccio, and B. Moore, "Nonlinear Evolution of Cosmological Structures in Warm Dark Matter Models," *Mon. Not. Roy. Astron. Soc.* **424** (2012) 684, arXiv:1112.0330 [astro-ph.CO].

[269] M. R. Lovell, "Towards a general parametrization of the warm dark matter halo mass function," *The Astrophysical Journal* **897** no. 2, (July, 2020) 147, arXiv:2003.01125 [astro-ph.CO].

[270] S. H. Suyu, P. J. Marshall, M. P. Hobson, and R. D. Blandford, "A bayesian analysis of regularised source inversions in gravitational lensing," *Mon. Not. Roy. Astron. Soc.* **371** (2006) 983–998, arXiv:astro-ph/0601493.

[271] S. Birrer and A. Amara, "Lenstronomy: multi-purpose gravitational lens modelling software package," *Physics of the Dark Universe* **22** (Dec., 2018) 189–201, arXiv:1803.09746 [astro-ph.CO].

[272] A. Galan, A. Peel, R. Joseph, F. Courbin, and J. L. Starck, "SLITronomy: towards a fully wavelet-based strong lensing inversion technique," *Astron. Astrophys.* **647** (2021) A176, arXiv:2012.02802 [astro-ph.GA].

[273] A. Adam, L. Perreault-Levasseur, and Y. Hezaveh, "Pixelated Reconstruction of Gravitational Lenses using Recurrent Inference Machines," in *39th International Conference on Machine Learning Conference*. 7, 2022. arXiv:2207.01073 [astro-ph.IM].

[274] P. Salucci, "The distribution of dark matter in galaxies," *The Astronomy and Astrophysics Review* **27** (Feb., 2019) , arXiv:1811.08843 [astro-ph.GA].

[275] E. Di Valentino *et al.*, "Snowmass2021 - Letter of interest cosmology intertwined I: Perspectives for the next decade," *Astropart. Phys.* **131** (2021) 102606, arXiv:2008.11283 [astro-ph.CO].

[276] B. Neichel, D. Mouillet, E. Gendron, C. Correia, J. F. Sauvage, and T. Fusco, "Overview of the european extremely large telescope and its instrument suite," arXiv:1812.06639 [astro-ph.IM].

[277] **Gaia** Collaboration, T. Prusti *et al.*, "The gaia mission," *Astronomy & Astrophysics* **595** (Nov, 2016) A1.

[278] J. Lazio, "The square kilometre array," arXiv:0910.0632 [astro-ph.IM].

[279] **CTA Consortium** Collaboration, J. Knödlseder, "The Cherenkov Telescope Array," in *16th Rencontres du Vietnam: Theory meeting experiment: Particle Astrophysics and Cosmology*. 4, 2020. arXiv:2004.09213 [astro-ph.IM].

[280] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, "Hybrid Monte Carlo," *Phys. Lett. B* **195** (1987) 216–222.

[281] J. Alsing, B. Wandelt, and S. Feeney, "Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology," *Mon. Not. Roy. Astron. Soc.* **477** no. 3, (2018) 2874–2885, arXiv:1801.01497 [astro-ph.CO].

[282] P. Lemos, M. Cranmer, M. Abidi, C. Hahn, M. Eickenberg, E. Massara, D. Yallup, and S. Ho, "Robust simulation-based inference in cosmology with Bayesian neural networks," *Mach. Learn. Sci. Tech.* **4** no. 1, (2023) 01LT01, arXiv:2207.08435 [astro-ph.CO].

[283] R. Legin, Y. Hezaveh, L. Perreault-Levasseur, and B. Wandelt, "A Framework for Obtaining Accurate Posteriors of Strong Gravitational Lensing Parameters with Flexible Priors and Implicit Likelihoods Using Density Estimation," *Astrophys. J.* **943** no. 1, (2023) 4, arXiv:2212.00044 [astro-ph.IM].

[284] X. Zhao, Y. Mao, and B. D. Wandelt, "Implicit Likelihood Inference of Reionization Parameters from the 21 cm Power Spectrum," *Astrophys. J.* **933** no. 2, (2022) 236, arXiv:2203.15734 [astro-ph.CO].

[285] S. Gagnon-Hartman, J. Ruan, and D. Haggard, "Debiasing standard siren inference of the Hubble constant with marginal neural ratio estimation," *Mon. Not. Roy. Astron. Soc.* **520** no. 1, (2023) 1–13, arXiv:2301.05241 [astro-ph.CO].

[286] B. Rhodes, K. Xu, and M. U. Gutmann, "Telescoping density-ratio estimation," in *34th Conference on Neural Information Processing Systems*. 06, 2020. arXiv:2006.12204 [stat.ML].

[287] M. Germain, K. Gregor, I. Murray, and H. Larochelle, "MADE: Masked Autoencoder for Distribution Estimation," in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach and D. Blei, eds., vol. 37, pp. 881–889. PMLR, Lille, France, 2, 2015. arXiv:1502.03509 [cs.LG].

[288] B. Uria, M.-A. Côté, K. Gregor, I. Murray, and H. Larochelle, "Neural autoregressive distribution estimation," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, G. Gordon, D. Dunson, and M. Dudík, eds., vol. 15, pp. 29–37. PMLR, 05, 2011. arXiv:1605.02226 [cs.LG].

[289] G. Papamakarios, T. Pavlakou, and I. Murray, "Masked Autoregressive Flow for Density Estimation," in *31st Conference on Neural Information Processing Systems*. 5, 2017. arXiv:1705.07057 [stat.ML].

[290] K. Scheutwinkel, W. Handley, C. Weniger, and E. de Lera Acedo, "PolySwyft: Likelihood-free nested sampling," 2023. expected October 2023.

[291] H. T. J. Bevins, W. J. Handley, P. Lemos, P. H. Sims, E. d. L. Acedo, A. Fialkov, and J. Alsing, "Marginal post-processing of Bayesian inference products with normalizing flows and kernel density estimators," *Mon. Not. Roy. Astron. Soc.* **526** no. 3, (2023) 4613–4626, arXiv:2205.12841 [astro-ph.IM].

[292] F. Lanusse, R. Mandelbaum, S. Ravanbakhsh, C.-L. Li, P. Freeman, and B. Poczos, "Deep generative models for galaxy image simulations," *Mon. Not. Roy. Astron. Soc.* **504** no. 4, (08, 2021) 5543–5555, 2008.03833 [astro-ph.IM].

[293] B. Uria, I. Murray, and H. Larochelle, "A deep and tractable density estimator," in *Proceedings of the 31st International Conference on Machine Learning*, E. P. Xing and T. Jebara, eds., vol. 32, pp. 467–475. PMLR, 10, 2014. arXiv:1310.1757 [stat.ML].

[294] W. J. Handley, M. P. Hobson, and A. N. Lasenby, "polychord: next-generation nested sampling," *Mon. Not. Roy. Astron. Soc.* **453** no. 4, (2015) 4385–4399, arXiv:1506.00171 [astro-ph.IM].

[295] J. Buchner, "Nested sampling methods," *Statistics Surveys* **17** no. none, (1, 2023) , arXiv:2101.09675 [stat.CO].

[296] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," arXiv:1512.03385 [cs.CV].

[297] D. Erkal and V. Belokurov, "Properties of Dark Subhaloes from Gaps in Tidal Streams," *Mon. Not. Roy. Astron. Soc.* **454** no. 4, (2015) 3542–3558, arXiv:1507.05625 [astro-ph.GA].

[298] N. Banik, G. Bertone, J. Bovy, and N. Bozorgnia, "Probing the nature of dark matter particles with stellar streams," *JCAP* **07** (2018) 061, arXiv:1804.04384 [astro-ph.CO].

[299] N. Banik and J. Bovy, "Effects of baryonic and dark matter substructure on the Pal 5 stream," *Mon. Not. Roy. Astron. Soc.* **484** no. 2, (2019) 2009–2020, arXiv:1809.09640 [astro-ph.GA].

[300] S. E. Koposov, H.-W. Rix, and D. W. Hogg, "Constraining the Milky Way potential with a 6-D phase-space map of the GD-1 stellar stream," *Astrophys. J.* **712** (2010) 260–273, arXiv:0907.1085 [astro-ph.GA].

[301] A. Bonaca, M. Geha, A. H. W. Küpper, J. Diemand, K. V. Johnston, and D. W. Hogg, "Milky Way Mass and Potential Recovery Using Tidal Streams in a Realistic Halo," *Astrophys. J.* **795** no. 1, (2014) 94, arXiv:1406.6063 [astro-ph.GA].

[302] N. C. Amorisco, F. A. Gòmez, S. Vegetti, and S. D. M. White, "Gaps in globular cluster streams: giant molecular clouds can cause them too," *Mon. Not. Roy. Astron. Soc.* **463** no. 1, (2016) L17–L21, arXiv:1606.02715 [astro-ph.GA].

[303] J. Bovy, A. Bahmanyar, T. K. Fritz, and N. Kallivayalil, "The shape of the inner Milky Way halo from observations of the Pal 5 and GD-1 stellar streams," *Astrophys. J.* **833** no. 1, (2016) 31, arXiv:1609.01298 [astro-ph.GA].

[304] K. Malhan, M. Valluri, K. Freese, and R. A. Ibata, "New Constraints on the Dark Matter Density Profiles of Dwarf Galaxies from Proper Motions of Globular Cluster Streams," *Astrophys. J. Lett.* **941** no. 2, (2022) L38, arXiv:2201.03571 [astro-ph.GA].

[305] **Gaia** Collaboration, A. G. A. Brown *et al.*, "Gaia Data Release 2: Summary of the contents and survey properties," *Astron. Astrophys.* **616** (2018) A1, arXiv:1804.09365 [astro-ph.GA].

[306] **Gaia** Collaboration, A. G. A. Brown *et al.*, "Gaia early data release 3: Summary of the contents and survey properties," *Astronomy & Astrophysics* **650** (2021) C3, arXiv:2012.01533 [astro-ph.GA].

[307] X. Cai, J. D. McEwen, and M. Pereyra, "Proximal nested sampling for high-dimensional bayesian model selection," *Statistics & Computing,* **32** (2002) 87, arXiv:2106.03646 [stat.ME].

[308] F. Feroz and M. P. Hobson, "Multimodal nested sampling: an efficient and robust alternative to MCMC methods for astronomical data analysis," *Mon. Not. Roy. Astron. Soc.* **384** (2008) 449, arXiv:0704.3704 [astro-ph].

[309] J. Alvey, M. Gerdes, and C. Weniger, "Albatross: Stellar stream parameter inference with neural ratio estimation." Astrophysics source code library, record ascl:2306.009, June, 2023.

[310] C. A. Correa, J. S. B. Wyithe, J. Schaye, and A. R. Duffy, "The accretion history of dark matter haloes – III. A physical model for the concentration–mass relation," *Mon. Not. Roy. Astron. Soc.* **452** no. 2, (2015) 1217–1232, arXiv:1502.00391 [astro-ph.CO].

[311] J. Alsing and B. Wandelt, "Generalized massive optimal data compression," *Mon. Not. Roy. Astron. Soc.* **476** no. 1, (2018) L60–L64, arXiv:1712.00012 [astro-ph.CO].

[312] C. Modi, S. Pandey, M. Ho, C. Hahn, B. Régaldo-Saint Blancard, and B. Wandelt, "Sensitivity Analysis of Simulation-Based Inference for Galaxy Clustering," arXiv:2309.15071 [astro-ph.CO].

[313] R. Barrué, P. Conde-Muíño, V. Dao, and R. Santos, "Simulation-based inference in the search for CP violation in leptonic WH production," *JHEP* **04** (2024) 014, arXiv:2308.02882 [hep-ph].

[314] L. Heinrich, S. Mishra-Sharma, C. Pollard, and P. Windischhofer, "Hierarchical Neural Simulation-Based Inference Over Event Ensembles," *Transactions on Machine Learning Research* (6, 2023) , arXiv:2306.12584 [stat.ML].

[315] K. Lin, M. von Wietersheim-Kramsta, B. Joachimi, and S. Feeney, "A simulation-based inference pipeline for cosmic shear with the Kilo-Degree Survey," *Mon. Not. Roy. Astron. Soc.* **524** no. 4, (2023) 6167–6180, arXiv:2212.04521.

[316] J. Poh, A. Samudre, A. Ćiprijanović, B. Nord, G. Khullar, D. Tanoglidis, and J. A. Frieman, "Strong Lensing Parameter Estimation on Ground-Based Imaging Data Using Simulation-Based Inference," in *36th Conference on Neural Information Processing Systems: Workshop on Machine Learning and the Physical Sciences.* 11, 2022. arXiv:2211.05836 [astro-ph.IM].

[317] E. Campeau-Poirier, L. Perreault-Levasseur, A. Coogan, and Y. Hezaveh, "Time Delay Cosmography with a Neural Ratio Estimator," in *40th International Conference on Machine Learning.* 9, 2023. arXiv:2309.16063 [astro-ph.IM].

[318] C. Hahn, M. Eickenberg, S. Ho, J. Hou, P. Lemos, E. Massara, C. Modi, A. Moradinezhad Dizgah, B. R.-S. Blancard, and M. M. Abidi, "SimBIG: mock challenge for a forward modeling approach to galaxy clustering," *JCAP* **04** (2023) 010, arXiv:2211.00660 [astro-ph.CO].

[319] Y. Brenier, U. Frisch, M. Henon, G. Loeper, S. Matarrese, R. Mohayaee, and A. Sobolevskii, "Reconstruction of the early universe as a convex optimization problem," *Mon. Not. Roy. Astron. Soc.* **346** (2003) 501–524, arXiv:astro-ph/0304214.

[320] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *International conference on machine learning*, pp. 1747–1756, PMLR. 2016. arXiv:1601.06759 [cs.CV].

[321] Y. Marnissi, D. Abboud, E. Chouzenoux, J.-C. Pesquet, M. El-Badaoui, and A. Benazza-Benyahia in *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5. 2019.

[322] R. A. C. Croft and E. Gaztanaga, "Reconstruction of cosmological density and velocity fields in the Lagrangian Zel'dovich approximation," *Mon. Not. Roy. Astron. Soc.* **285** (1997) 793, arXiv:astro-ph/9602100.

[323] U. Frisch, S. Matarrese, R. Mohayaee, and A. Sobolevski, "A Reconstruction of the initial conditions of the universe by optimal mass transportation," *Nature* **417** (2002) 260–262, arXiv:astro-ph/0109483.

[324] M. Gramann, "An Improved Reconstruction Method for Cosmological Density Fields," *The Astrophysical Journal* **405** (Mar., 1993) 449.

[325] A. Nusser and A. Dekel, "Tracing Large-Scale Fluctuations Back in Time," *The Astrophysical Journal* **391** (June, 1992) 443.

[326] P. J. E. Peebles, *The large-scale structure of the universe*, vol. 98. Princeton university press, 2020.

[327] D. H. Weinberg, "Reconstructing primordial density fluctuations. I - Method," *Mon. Not. Roy. Astron. Soc.* **254** (Jan., 1992) 315–342.

[328] Y. Li, C. Modi, D. Jamieson, Y. Zhang, L. Lu, Y. Feng, F. Lanusse, and L. Greengard, "Differentiable Cosmological Simulation with the Adjoint Method," *Astrophys. J. Suppl.* **270** no. 2, (2024) 36, arXiv:2211.09815 [astro-ph.IM].

[329] C. Modi, F. Lanusse, and U. Seljak, "FlowPM: Distributed TensorFlow implementation of the FastPM cosmological N-body solver," *Astron. Comput.* **37** (2021) 100505, arXiv:2010.11847 [astro-ph.CO].

[330] C. J. Shallue and D. J. Eisenstein, "Reconstructing cosmological initial conditions from late-time structure with convolutional neural networks," *Mon. Not. Roy. Astron. Soc.* **520** no. 4, (2023) 6256–6267, arXiv:2207.12511 [astro-ph.CO].

[331] X. Chen, F. Zhu, S. Gaines, and N. Padmanabhan, "Effective cosmic density field reconstruction with convolutional neural network," *Mon. Not. Roy. Astron. Soc.* **523**

no. 4, (2023) 6272–6281, arXiv:2306.10538 [astro-ph.CO].

[332] C. Modi, Y. Li, and D. Blei, "Reconstructing the universe with variational self-boosted sampling," *JCAP* **03** (2023) 059, arXiv:2206.15433 [astro-ph.IM].

[333] C. Modi, F. Lanusse, U. Seljak, D. N. Spergel, and L. Perreault-Levasseur, "CosmicRIM : Reconstructing Early Universe by Combining Differentiable Simulations with Recurrent Inference Machines," arXiv:2104.12864 [astro-ph.CO].

[334] R. Legin, M. Ho, P. Lemos, L. Perreault-Levasseur, S. Ho, Y. Hezaveh, and B. Wandelt, "Posterior sampling of the initial conditions of the universe from non-linear large scale structures using score-based generative models," *Mon. Not. Roy. Astron. Soc.* **527** no. 1, (2023) L173–L178, arXiv:2304.03788 [astro-ph.CO].

[335] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," arXiv:1502.03167 [cs.LG].

[336] C. Holmes and S. Walker, "Assigning a value to a power likelihood in a general bayesian model," *Biometrika* no. 2, (2017) 497–503, arXiv:1701.08515 [stat.ME].

[337] J. Jasche and G. Lavaux, "Physical bayesian modelling of the non-linear matter distribution: new insights into the nearby universe," *Astronomy & Astrophysics* **625** (2019) A64, arXiv:1806.11117 [astro-ph.CO].

[338] M. Vono, N. Dobigeon, and P. Chainais, "High-dimensional gaussian sampling: a review and a unifying approach based on a stochastic proximal point algorithm," *SIAM Review* **64** no. 1, (10, 2020) 3–56, arXiv:2010.01510 [stat.CO].

[339] F. R. Bouchet, R. Juszkiewicz, S. Colombi, and R. Pellat, "Weakly Nonlinear Gravitational Instability for Arbitrary Omega," *Astrophysical Journal Letters* **394** (July, 1992) L5.

[340] T. Buchert and J. Ehlers, "Lagrangian theory of gravitational instability of Friedman-Lemaitre cosmologies – second-order approach: an improved model for non-linear clustering," *Mon. Not. Roy. Astron. Soc.* **264** (Sept., 1993) 375–387.

[341] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241, Springer. Springer International Publishing, Cham, 2015. arXiv:1505.04597 [cs.CV].

[342] A. Graps, "An introduction to wavelets," *IEEE computational science and engineering* **2** no. 2, (1995) 50–61.

[343] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," *Advances in neural information processing systems* **28** (2015) , arXiv:1506.03099 [cs.LG].

[344] C. Rampf, S. O. Schobesberger, and O. Hahn, "Analytical growth functions for cosmic structures in a ΛCDM Universe," *Mon. Not. Roy. Astron. Soc.* **516** no. 2, (2022) 2840–2850, arXiv:2205.11347 [astro-ph.CO].

[345] R. W. Hockney and J. W. Eastwood, *Computer Simulation Using Particles (1st ed.)*. CRC Press, 1988.

[346] A. Weltman *et al.*, "Fundamental physics with the Square Kilometre Array," *Publ. Astron. Soc. Austral.* **37** (2020) e002, arXiv:1810.02680 [astro-ph.CO].

[347] **CTA Consortium** Collaboration, B. S. Acharya *et al.*, *Science with the Cherenkov Telescope Array*. WSP, 11, 2018. arXiv:1709.07997 [astro-ph.IM].

[348] B. Panes, C. Eckner, L. Hendriks, S. Caron, K. Dijkstra, G. Jóhannesson, R. R. de Austri, and G. Zaharijas, "Identification of point sources in gamma rays using U-shaped convolutional neural networks and a data challenge," *Astron. Astrophys.* **656** (2021) A62, arXiv:2103.11068 [astro-ph.HE].

[349] S. Caron, G. A. Gómez-Vargas, L. Hendriks, and R. Ruiz de Austri, "Analyzing γ-rays of the Galactic Center with Deep Learning," *JCAP* **05** (2018) 058, arXiv:1708.06706 [astro-ph.HE].

[350] F. List, N. L. Rodd, and G. F. Lewis, "Extracting the Galactic Center excess' source-count distribution with neural nets," *Phys. Rev. D* **104** no. 12, (2021) 123022, arXiv:2107.09070 [astro-ph.HE].

[351] A. Vafaei Sadr, E. E. Vos, B. A. Bassett, Z. Hosenie, N. Oozeer, and M. Lochner, "DeepSource: Point Source Detection using Deep Learning," *Mon. Not. Roy. Astron. Soc.* **484** no. 2, (2019) 2793–2806, arXiv:1807.02701 [astro-ph.IM].

[352] S. Rezaei, J. P. McKean, M. Biehl, and A. Javadpour, "Decoras: detection and characterization of radio-astronomical sources using deep learning," *Mon. Not. Roy. Astron. Soc.* **510** no. 4, (Dec., 2021) 5891–5907, arXiv:2109.09077 [astro-ph.IM].

[353] V. Lukic, F. D. Gasperin, and M. Brüggen, "Convosource: Radio-astronomical source-finding with convolutional neural networks," *Galaxies* (10, 2019) , arXiv:1910.03631 [astro-ph.IM].

[354] D. Tilley, C. W. Cleghorn, K. Thorat, and R. Deane, "Point proposal network: Accelerating point source detection through deep learning," *2021 IEEE Symposium Series on Computational Intelligence (SSCI), 2021, pp. 1-8* (08, 2020) , arXiv:2008.02093 [cs.CV].

[355] L. Bonavera, S. L. S. Gomez, J. González-Nuevo, M. M. Cueli, J. D. Santos, M. L. Sanchez, R. Muñiz, and F. J. de Cos, "Point source detection with fully-convolutional networks: Performance in realistic simulations," *Astronomy & Astrophysics* **648** (2021) A50, arXiv:1911.11826 [astro-ph.GA].

[356] H. Cholakkal, G. Sun, F. S. Khan, and L. Shao, "Object counting and instance segmentation with image-level supervision," arXiv:1903.02494 [cs.CV].

[357] V. Ranjan, U. Sharma, T. Nguyen, and M. Hoai, "Learning to count everything," arXiv:2104.08391 [cs.CV].

[358] E. Kilic and S. Ozturk, "An accurate car counting in aerial images based on convolutional neural networks," *Journal of Ambient Intelligence and Humanized*

Computing (Jul, 2021) .

[359] P. M. S. Blackett, "The Rutherford Memorial Lecture, 1958," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **251** no. 1266, (1959) 293–305.

Summary

The aim of physics is to uncover the underlying principles governing natural phenomena. These laws are articulated through theoretical models, often in mathematical terms (such as Newton's law of gravitation), and must be tested against experimental or observational evidence (for instance, measuring how long it takes for an apple to fall). A crucial aspect of physics involves quantitatively comparing these theoretical models with empirical data and measurements through statistical analysis.

The primary goal of this thesis is to enable new physics discoveries by addressing the statistical and computational challenges that arise in the fields of astrophysics and cosmology. Traditional methods in modern astrophysical data analysis are sampling-based inference techniques like Markov-chain Monte Carlo and nested sampling methods. However, these approaches frequently rely on approximate likelihoods and suffer from a significant drawback: the time required to achieve convergence scales poorly with the dimensionality of the explored parameter space. To overcome these limitations, this thesis contributes to develop and establish an alternative approach based on novel simulation-based inference (SBI) techniques, that have seen a remarkable development in recent years.

We began in Chapter 2 by outlining the landscape of SBI approaches, including traditional ones, such as approximate Bayesian computation. We then detailed the various neural SBI algorithms, neural posterior estimation and neural likelihood estimation, and especially focused on truncated marginal neural ratio estimation (TMNRE), the main inference algorithm employed throughout this thesis. TMNRE builds on three key ingredients: neural ratio estimation via classification, marginalization, and prior truncation. We then emphasized its different advantages and pitfalls with respect to the others SBI methods, and discussed testing strategies.

We first applied TMNRE in Chapter 3, to the analysis of strong lensing images as a dark matter probe. Starting with an overview of strong lensing and its potential to probe dark matter substructures, we discussed the modeling complexities involved, including lens and source parameter uncertainties, and how to model populations of substructures from different dark matter scenarios. We developed a SBI pipeline

to infer the cutoff mass in the dark matter halo mass function directly from simulated observations. Our results demonstrated that TMNRE enables precise marginal and targeted inference, overcoming traditional computational challenges. We further showed the application of hierarchical inference to extract the dark matter cutoff mass signal from a dataset of lenses, paving the way for future advancements in dark matter characterization using strong lensing observations.

We then extended the TMNRE framework in Chapter 4, where we proposed two new technical building blocks. First we developed autoregressive ratio estimation with the aim to robustly estimate correlated high-dimensional posteriors. Second, we proposed a slice-based nested sampling algorithm to efficiently draw both posterior samples and constrained prior samples from ratio estimators, the latter being instrumental for sequential inference.

In Chapter 5, we tackled the inverse problem that goes from non-linear, non-local mappings of late-time density fields to Gaussian cosmological initial conditions. To this end, we employed autoregressive Gaussian likelihood estimation to model the conditional dependencies between pixels in the density field. Posterior sampling is achieved through a Gibbs sampling algorithm based on exact data augmentation, ensuring efficient exploration of high-dimensional parameter spaces. The proposed approach combines computational efficiency with applicability to generic, non-differentiable forward simulators, making it suitable for broader astrophysical and cosmological data analysis tasks.

Finally, in Chapter 6, we turned towards the problem of point source detection and population parameters inference in sky-maps. We developed a highly interpretable (since it resembles components of traditional survey analysis workflows) SBI framework that allows, for the first time to the authors' knowledge, to perform consistently point source detection and population parameters inference, from both detected and sub-threshold sources. This was possible by defining source detection as a novel and high-dimensional form of prior truncation to incorporate detected sources into the simulation model.

Overall, this thesis aimed to highlight the potential of SBI for astrophysical data analysis, positioning this framework as an essential part of the modern physics data analysis toolkit.

Samenvatting

Het doel van de natuurkunde is de onderliggende principes te onthullen die natuurlijke fenomenen beheersen. Deze wetten worden uitgedrukt via theoretische modellen, vaak in wiskundige termen (zoals de zwaartekrachtswet van Newton), en moeten worden getoetst aan experimenteel of observationeel bewijs (bijvoorbeeld door te meten hoe lang het duurt voordat een appel valt). Een cruciaal aspect van de natuurkunde is het kwantitatief vergelijken van deze theoretische modellen met empirische gegevens en metingen door middel van statistische analyse.

Het primaire doel van dit proefschrift is nieuwe natuurkundige ontdekkingen mogelijk te maken door de statistische en computationele uitdagingen aan te pakken die zich voordoen in de vakgebieden van de astrofysica en kosmologie. Traditionele methoden in moderne astrofysische data-analyse zijn op steekproeven gebaseerde inferentietechnieken zoals Markov-chain Monte Carlo en geneste sampling-methoden. Deze strategieën vertrouwen echter vaak op benaderende likelihoods en hebben een belangrijk nadeel: de tijd die nodig is om tot convergentie te komen, schaalt slecht met de dimensionaliteit van de verkende parameterruimte. Om deze beperkingen te overwinnen, draagt dit proefschrift bij aan de ontwikkeling en vestiging van een alternatieve strategie op basis van nieuwe simulation-based inference (SBI) technieken, die de afgelopen jaren een significante ontwikkeling hebben doorgemaakt.

We begonnen in Hoofdstuk 2 met het schetsen van het landschap van SBI-strategieën, waaronder traditionele methoden zoals approximate Bayesian computation. Vervolgens hebben we de verschillende neurale SBI-algoritmen beschreven, zoals neurale posterior schatting en neurale likelihood schatting, waarbij we ons met name hebben gericht op truncated marginal neural ratio estimation (TMNRE), het voornaamste inferentie-algoritme dat in dit proefschrift wordt gebruikt. TMNRE is gebaseerd op drie belangrijke ingrediënten: neurale ratio schatting via classificatie, marginalisatie, en prior-truncatie. We benadrukten vervolgens de verschillende voordelen en valkuilen ten opzichte van de andere SBI-methoden, en bespraken teststrategieën.

We hebben TMNRE voor het eerst toegepast in Hoofdstuk 3, op de analyse van

sterke-lensbeelden als probe voor donkere materie. Beginnend met een overzicht van sterke-lenswerking en het potentieel ervan om substructuren van donkere materie te onderzoeken, bespraken we de complexiteiten van de modellering die hierbij betrokken zijn, waaronder onzekerheden in de lens- en bronparameters, en hoe populaties van substructuren uit verschillende scenario's voor donkere materie gemodelleerd kunnen worden. We hebben een SBI-pijplijn ontwikkeld om de cutoff-massa in de massafunctie van donkerematerie halos rechtstreeks af te leiden uit gesimuleerde waarnemingen. Onze resultaten toonden aan dat TMNRE nauwkeurige marginale en gerichte inferentie mogelijk maakt en daarmee traditionele computationele uitdagingen overwint. We lieten verder zien hoe hiërarchische inferentie kan worden toegepast om de cutoff-massa van donkere materie uit een dataset van lenzen te extraheren, wat de weg vrijmaakt voor toekomstige verbeteringen in de karakterisering van donkere materie met behulp van sterke-lenswaarnemingen.

Vervolgens hebben we het TMNRE-framework uitgebreid in Hoofdstuk 4, waar we twee nieuwe technische bouwstenen voorstelden. Ten eerste ontwikkelden we autoregressieve ratio schatting met als doel om op robuuste wijze gecorreleerde hoog-dimensionale posteriors te schatten. Ten tweede stelden we een op slices gebaseerd genest sampling-algoritme voor om efficiënt zowel posterior samples als beperkte prior samples te genereren uit ratio-schatters, waarbij dit laatste instrumenteel is voor sequentiële inferentie.

In Hoofdstuk 5 hebben we het inverse probleem aangepakt dat zich richt op de overgang van niet-lineaire, niet-lokale beschrijvingen van dichtheidsvelden in het late universum naar Gaussische kosmologische beginvoorwaarden. Hiervoor hebben we autoregressieve Gaussische likelihood-schatting gebruikt om de voorwaardelijke afhankelijkheden tussen pixels in het dichtheidsveld te modelleren. Posterior sampling wordt uitgevoerd via een Gibbs-sampling-algoritme op basis van exacte data-augmentatie, wat zorgt voor efficiënte verkenning van hoog-dimensionale parameterruimten. De voorgestelde aanpak combineert computationele efficiëntie met toepasbaarheid op generieke, niet-differentieerbare forward simulators, waardoor deze geschikt is voor bredere astrofysische en kosmologische data-analysetaken.

Tot slot hebben we in Hoofdstuk 6 het probleem van puntbron-detectie en populatieparameterinferentie in hemelkaarten behandeld. We ontwikkelden een goed interpreteerbaar SBI-framework (omdat het lijkt op componenten van traditionele survey-analyseworkflows) dat, voor zover de auteurs weten, voor het eerst consistente puntbron-detectie en populatieparameterinferentie mogelijk maakt, zowel voor gedetecteerde als sub-drempelbronnen. Dit werd mogelijk gemaakt door brondetectie te definiëren als een nieuwe en hoog-dimensionale vorm van prior-truncatie om gedetecteerde bronnen op te nemen in het simulatiemodel.

Over het geheel genomen was het doel van dit proefschrift om het potentieel van

SBI voor astrofysische data-analyse te benadrukken, en dit framework te positioneren als een essentieel onderdeel van de moderne toolkit voor natuurkundige data-analyse.

Vertaald door Dion Noordhuis.

Acknowledgements

First and foremost I'd like to thank my supervisor Christoph, who taught me how to do research and how to navigate the hectic world of academia. When I started working with you I did not even know how to run a Python script, let alone training a neural network. And here I am, four years later, with a thesis on machine learning for astrophysical data analysis. Thank you for your trust, patient guidance, and complete support. I'm especially grateful for the freedom you gave me to pursue and explore various projects, often inspired by the many ideas you generate like a volcano.

Being at GRAPPA over the last four years has been an amazing experience. I'd like to thank all GRAPPA faculty — Gianfranco, Samaya, Philipp, Shin'ichiro, Jacco, Ben, and Sera — for fostering a thriving workplace with your availability, openness, and enthusiasm. A special thank goes to my independent supervisor, Wouter Waalewijn, who has always ensured that I felt comfortable and free to talk. No institute can succeed without the tireless effort of the administration staff. In particular, Jiřina Šálková, GRAPPA is lucky to have you and I'm hugely grateful for your availability and help in navigating me through life at a scientific institute.

All the past and present GRAPPA PhD students deserve a special thank you for being such wonderful office mates: Ariane, Banafsheh, Ben, Uddipta, Gimmy, Dion, Fabian, Mathis, Oleg, Youyou, Alexandra, Kate, Huyfang, and more recently Pravita and Joy. Gimmy, abbiamo iniziato il dottorato lo stesso giorno, lo finiamo insieme. Mi mancheranno le nostre passeggiate senza meta (ma con tappa gelato) e i giri in bici in cui più di una volta ho confuso destra e sinistra... ma non il tuo "risotto" agli spinaci! Non vedo l'ora di organizzare le prossime camminate, possibilmente in salita. From the moment I first entered the GRAPPA PhD office, Uddipta welcomed me with his characteristic warmth. Your kindness never fades — thank you for all the great times we've shared and your adventurous stories. Fabian, thank you for making our lunch breaks so much more engaging with insightful discussions. I'll greatly miss our chats. Special thanks goes to my paranymph Dion for translating my English Summary to the Dutch Samenvatting, and for the fun times we shared at the GGI school.

To the current and past postdocs at GRAPPA — Sam, Pippa, Camila, Oscar, Pablo, James, Guillermo, Malte — thank you for all your advice, feedback (especially during application season!), and the postdoc-wisdom you have shared. And especially thank you for all the fun memories and the pub-quizzes! In particular, Camila, you have been a wonderful daily supervisor. Thank you for guiding me through the analysis of cosmological simulations. James, since you arrived, you have been a constant source of guidance for me. You taught me a lot about science and what it means to be a scientist through our numerous discussions, for which I'm very grateful. Your insightful ability to communicate - and willingness to argue - has been eye-opening. Thank you for your time and dedication to our research; I look forward to what the future holds.

To my first collaborators, Adam and Kosio, thank you a lot for introducing me to the magical world of lensing and helping me in my first steps in learning how to make a machine learn. I have learnt a great deal from both of you.

During my PhD, I had the pleasure of meeting and collaborating with many other great people. Special thanks to Florian, for introducing me to cosmology and for a fun coding week in Annecy, and to Christopher, for introducing me to the gamma-ray sky analysis and its wonders. Another big thanks goes to Francesca, who was a wonderful host at LAPTh in Annecy.

Outside of academia, there are many people that have made my life full, and contributed just as much through their presence.

Fra, dai giorni in cui vivevamo nel mio studio senza mobili, a quelli sul tuo terrazzo con vista giraffe, alle ultime passeggiate lungo i canali, sei stata il mio pezzo di casa fuori casa. Sono molto felice di aver potuto condividere la maggior parte di questi anni ad Amsterdam con te. Alle nostre 'Orchidee' — Cha, Eri, Sof, Agne ed Ely — grazie per esserci venute a trovare portando con voi altri pezzi di casa.

Marcelo and Celina, you were the first ones to welcome me in Amsterdam, in Esnoga, and in your house. Thank you for always hosting me during many Hagim. Emmanuel, thank you for our Amsterdam strolls and inspiring conversations. Though we don't see each other often, I value how close we've stayed since arriving in Amsterdam around the same time.

La 'LaMamma', credo che l'articolo — parte integrante di come mi rivolgo a te — dica molte cose. Aggiungo un grazie per avermi insegnato l'importanza dell'attenzione e della cura, in ogni loro aspetto e manifestazione. Papà, se ho difficoltà a compiere scelte è merito tuo, che mi hai mostrato e sempre offerto quante possibilità esistono e la libertà di coglierle. Gio, grazie di aver badato a casa e specialmente alla Sbisu.

Una presenza costante nonostante la distanza — Gio, non vedo l'ora di continuare a camminare insieme.

Noemi,
October 2024