Lecture Notes in Physics

Nima Arkani-Hamed · Mathieu Giroux · Holmfridur Sigridar Hannesdottir · Sebastian Mizera · Celina Pasiecznik *Editors*

Records from the *S*-Matrix Marathon

Selected Topics on Scattering Amplitudes

Springer

Lecture Notes in Physics

Founding Editors

Wolf Beiglböck

Jürgen Ehlers

Klaus Hepp

Hans-Arwed Weidenmüller

Volume 1041

Series Editors

Roberta Citro, Salerno, Italy

Peter Hänggi, Augsburg, Germany

Betti Hartmann, London, UK

Morten Hjorth-Jensen, Oslo, Norway

Maciej Lewenstein, Barcelona, Spain

Satya N. Majumdar, Orsay, France

Luciano Rezzolla, Frankfurt am Main, Germany

Angel Rubio, Hamburg, Germany

Wolfgang Schleich, Ulm, Germany

Stefan Theisen, Potsdam, Germany

James D. Wells, Ann Arbor, MI, USA

Gary P. Zank, Huntsville, AL, USA

The series Lecture Notes in Physics (LNP), founded in 1969, reports new developments in physics research and teaching - quickly and informally, but with a high quality and the explicit aim to summarize and communicate current knowledge in an accessible way. Books published in this series are conceived as bridging material between advanced graduate textbooks and the forefront of research and to serve three purposes:

- to be a compact and modern up-to-date source of reference on a well-defined topic;
- to serve as an accessible introduction to the field to postgraduate students and non-specialist researchers from related areas;
- to be a source of advanced teaching material for specialized seminars, courses and schools.

Both monographs and multi-author volumes will be considered for publication. Edited volumes should however consist of a very limited number of contributions only. Proceedings will not be considered for LNP.

Volumes published in LNP are disseminated both in print and in electronic formats, the electronic archive being available at springerlink.com. The series content is indexed, abstracted and referenced by many abstracting and information services, bibliographic networks, subscription agencies, library networks, and consortia.

Proposals should be sent to a member of the Editorial Board, or directly to the responsible editor at Springer:

Dr Lisa Scalone
lisa.scalone@springernature.com

Nima Arkani-Hamed • Mathieu Giroux •
Holmfridur Sigridar Hannesdottir •
Sebastian Mizera • Celina Pasiecznik
Editors

Records from the S-Matrix Marathon

Selected Topics on Scattering Amplitudes

Springer

Editors
Nima Arkani-Hamed
Institute for Advanced Study
Princeton, NJ, USA

Holmfridur Sigridar Hannesdottir
Institute for Advanced Study
Princeton, NJ, USA

Celina Pasiecznik
Department of Physics
McGill University
Montreal, QC, Canada

Mathieu Giroux
Department of Physics
McGill University
Montreal, QC, Canada

Sebastian Mizera
Department of Physics
Princeton University
Princeton, NJ, USA

ISSN 0075-8450 ISSN 1616-6361 (electronic)
Lecture Notes in Physics
ISBN 978-3-031-90351-9 ISBN 978-3-031-90352-6 (eBook)
https://doi.org/10.1007/978-3-031-90352-6

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Abstract

Recent years have seen enormous progress in understanding foundations of the S-matrix theory from different points of view, ranging from their analytic properties and infrared divergences, through applications to gravitational-wave observables and cosmological correlators, all the way to lattice simulations and the phenomenology of strongly coupled QCD. All these approaches are converging into a unified picture for understanding particle interactions. They were discussed during the *S-Matrix Marathon*: a workshop hosted at the Institute for Advanced Study in Princeton on 11–22 March 2024. These lecture notes are the records from the marathon. Their purpose is to provide a pedagogical introduction to the unfolding ideas surrounding the S-matrix theory and to shed light on the emerging directions.

Preface

You are reading the records from the *S-Matrix Marathon*, a workshop held at the Institute for Advanced Study in Princeton, New Jersey, during 11–22 March 2024. Motivation for organizing this workshop was many-fold. Recent years have seen an outburst of activity in studying the analytic properties of the S-matrix and related objects from different angles. In particular, the historically challenging multi-particle case has seen a lot of progress. Moreover, the field has gradually started establishing far-reaching connections between scattering amplitudes and observables in curved spaces, on time-folded contours, or using non-perturbative techniques to understand experimental data or lattice simulations. Seeing more of such connections on the horizon, three of us decided to host a workshop designed to learn the techniques used across different approaches to scattering and foster new collaborations.

The aim of this workshop was to prioritize depth rather than breadth of the topics covered. For this reason, we gave each speaker an entirely free hand at organizing one full day of activities. In the invitation email, we emphasized that they can do whatever they see fit: invite supporting speakers, give 10 hours of lectures, moderate informal discussions, or go to the beach. Alas, no one chose the last option, though some floated it. In the end, we had a true marathon of lectures. After hours, the participants joined in on some of the many events organized at the Institute at the same time, including a music performance by the Ensemble 132 and a public talk by David Kaiser on the science and history of the Global Positioning System.

We were particularly impressed by the pedagogy of the marathon lecturers, which prompted us to record their content as lecture notes. You are reading them now. We believe they will be an important resource for the community. Most of the records were written almost entirely by us, the undersigned editors. They were later sent to the speakers for improvements and corrections. All blame for errors and typographical mistakes rests on us. We are also guilty of all marathon-related puns included in this text, for which we apologize in advance.

Let us give an outline of the topics covered throughout this workshop and Records. The subject largely evolves around understanding the analytic properties of the S-matrix, so it seems natural to start by reviewing some facts from complex analysis in several complex variables. Sean Curry and Jiří Lebl started off the marathon by presenting some of the "tasty bits" of this subject in Part 1. The topic

of computing domains of holomorphy resonates particularly well with some of the later lectures.

One of the main challenges in the S-matrix theory is extracting non-perturbative information. A rare window on this problem is provided by lattice field theory, which—largely thanks to understanding analyticity properties of the amplitudes—has matured enough to be able to compute $3 \to 3$ matrix elements of pions. Max Hansen provided a pedagogical introduction to this topic, which is summarized in Part 2. It reviews the use of the Bethe–Salpeter equation in this setup, which later reappears in the context of gravitational scattering.

Dispersion relations provide sharp statements encoding the analytic properties of scattering amplitudes and decay rates. They can be used to extract and constrain the information from experimental data. During the workshop, Emilie Passemar gave an outline of this fascinating subject connecting theory and practice. Dispersion relations also appear later on in the context of the S-matrix bootstrap.

One of the exciting aspects of quantum field theory is that it allows us to play around with the concept of time. As it turns out, in certain situations it pays off to think about it as not just running forwards but instead twisting and bending back and forth. The theory behind these observables is the Schwinger–Keldysh formalism. Felix Haehl and Mukund Rangamani gave a pedagogical introduction to this branch of physics. Their tour de force exposition, starting from thermal physics and ending on the relations to chaos and gravitational path integrals, is given in Part 3. The Schwinger–Keldysh formalism reappears in later lectures on cosmological observables and gravitational waveforms.

Rigorous foundations for scattering theory, at least in the mass-gapped case, have been developed in the last century. In view of the applications to the S-matrix bootstrap, it is important to revisit such constructions and start asking what could be some alternative starting points for the rigorous definition of the S-matrix. During the workshop, Aditya Hebbar and Balt van Rees reviewed the Haag–Ruelle scattering theory, gave a summary of the proposals for defining scattering amplitudes using the AdS/CFT dictionary, and finally showcased some applications of the S-matrix bootstrap to constraining observables in gauge theories.

Another approach to curved backgrounds is to study observables in cosmological backgrounds and their connection to the flat-space S-matrix. The second week of the marathon started with Paolo Benincasa and Francisco Vazão, who gave a lucid introduction to this much-studied topic, summarized in Part 4. They explain what is known about the connections between the cosmological wave function and scattering amplitudes, as well as how tools translate from one topic to another.

One way in which causality is encoded in the S-matrix is through the property called crossing, which relates scattering amplitudes to their result after analytic continuation from positive to negative energies and vice versa. In Part 5, Simon Caron-Huot and Mathieu Giroux review new advances in this subject and explain how crossing relates time-ordered scattering amplitudes to a zoo of other observables where the time ordering is relaxed. One application is to the computation of expectation values of operators.

Another connection to analytically continuing time, this time in cosmology, has been presented by Gordon Lee and Enrico Pajer and recorded in Part 6. After reviewing aspects of cosmological correlators, their analytic properties and cutting rules, they discuss new computational tools for practical calculations.

Recent years have seen an outburst of activity in studying the emergence of classical physics from scattering amplitudes, in particular with the view of applications to gravitational-wave physics. The penultimate day of the marathon was a medley of results on this topic. Miguel Correia and Giulia Isabella gave a pedagogical introduction to the eikonal approximation and showed how some of the classic results in Quantum Electrodynamics and General Relativity can be recovered using the diagrammatic approach. Three short talks on the related topics brought us through the finish line: Zihan Zhou talked about some astrophysical application of worldline effective field theory, Hofie Hannesdottir explained the role of radiation in gravitational observables, and Anna Wolz showed results on the 3-body problem in General Relativity.

The last day of the workshop featured *sprint talks* by some of the participants who volunteered to speak: Carolina Figueiredo, Alessandro Podo, Andrzej Pokraka, Francesco Serra, and Cristian Vergu. We also had a trip to Albert Einstein's house and a visit to the Institute's rare book collection and the archives, where the marathon participants had a chance to view the first printed edition of Euclid's Elements and try on Einstein's stylish glasses.

Princeton, NJ, USA — Nima Arkani-Hamed
Montreal, QC, Canada — Mathieu Giroux
Princeton, NJ, USA — Holmfridur Sigridar Hannesdottir
Princeton, NJ, USA — Sebastian Mizera
Montreal, QC, Canada — Celina Pasiecznik

Acknowledgments

S.C.H., M.G., and C.P.'s work is supported in part by the National Science and Engineering Council of Canada (NSERC) and the Canada Research Chair program (reference number CRC-2022-00421). S.C.H. is additionally supported by a Simons Fellowship in Theoretical Physics and by the Simons Collaboration on the Nonperturbative Bootstrap, while C.P. is further supported by the Walter C. Sumner Memorial Fellowship. M.C. is supported by the Simons Collaboration on the Nonperturbative Bootstrap. S.C. and J.L. would like to thank the organizers for the wonderful conference (and for bringing in mathematicians); S.C. is supported in part by the Simons Foundation (grant MPS-TSM-00002876) and J.L. in part by the Simons Foundation collaboration grant 710294. H.S.H. gratefully acknowledges funding provided by the J. Robert Oppenheimer Endowed Fund of the Institute for Advanced Study, and S.M. gratefully acknowledges funding provided by the Sivian Fund and the Roger Dashen Member Fund at the Institute for Advanced Study. G.I. is supported by the US Department of Energy under award number DE-SC0024224, the Sloan Foundation, and the Mani L. Bhaumik Institute for Theoretical Physics, while A.M.W. is supported by the US Department of Energy under award number DE-SC0024224, the Alfred P. Sloan Foundation, and the NSF Graduate Research Fellowship under grant no. DGE-2034835, with thanks to the Mani L. Bhaumik Institute for Theoretical Physics. E.P. and M.H.G.L.'s work has been supported by the STFC consolidated grants ST/X001113/1, ST/T000694/1, ST/X000664/1, and EP/V048422/1. P.B. thanks Dieter Lüst and Gia Dvali for making his participation possible and acknowledges the hospitality of the Galician Institute for High Energy Physics (IGFAE) and the University of Santiago de Compostela. This material is based upon work supported by the US Department of Energy, Office of Science, Office of High Energy Physics under Award Number DE-SC0009988.

The S-Matrix Marathon workshop was sponsored by the Institute for Advanced Study and the Carl P. Feinberg Program in Cross-Disciplinary Innovation.

Contents

Tasty Bits of Several Complex Variables

1

Sean Curry and Jiří Lebl

Abstract

We will cover the basics of several complex variables in 4 lectures: Basic properties of holomorphic functions in several variables, the notion of pseudoconvexity, CR functions and CR geometry, and the $\bar{\partial}$-problem. The main underlying idea is to connect various characterizations of domains of holomorphy, that is, the natural domains of definition for holomorphic functions. In the process we will connect the function theory on the domain to geometric properties of the boundary, and discuss the relationship between the boundary values and the functions themselves and extension of holomorphic functions from subspaces.

These lecture notes roughly correspond to the first 4 chapters of the book of the same name (J. Lebl [1], *Tasty bits of several complex variables* Independently published: https://www.jirka.org/scv/, (2023)).

S. Curry · J. Lebl (✉)
Departemento pri Matematiko, Oklahoma Ŝtata Universitato, Stillwater, OK, USA
e-mail: lebl@okstate.edu; lebl@okstate.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

N. Arkani-Hamed et al. (eds.), *Records from the S-Matrix Marathon*,
Lecture Notes in Physics 1041, https://doi.org/10.1007/978-3-031-90352-6_1

1.1 Holomorphic Functions in Several Variables

Jiří Lebl

1.1.1 Onto Several Variables

These lecture notes roughly correspond to the first 4 chapters of the book [1] of the same name.

Let $\mathbb{C}^n$ be the complex Euclidean space with coordinates $z = (z_1, z_2, \ldots, z_n) \in \mathbb{C}^n$ and it will be useful to treat it as two copies of the real space, $\mathbb{C}^n \cong \mathbb{R}^n \times \mathbb{R}^n = \mathbb{R}^{2n}$ with

$$z = x + iy, \qquad \bar{z} = x - iy, \qquad x, y \in \mathbb{R}^n, \qquad i = \sqrt{-1}. \tag{1.1.1}$$

We call z the *holomorphic coordinates* and $\bar{z}$ *antiholomorphic coordinates*. Let us define a *polydisc* $\Delta_\rho(a)$ with *polyradius* $\rho = (\rho_1, \rho_2, \ldots, \rho_n)$ and *center* $a \in \mathbb{C}^n$ as

$$\Delta_\rho(a) \stackrel{\text{def}}{=} \left\{ z \in \mathbb{C}^n : |z_k - a_k| < \rho_k \text{ for } k = 1, 2, \ldots, n \right\}. \tag{1.1.2}$$

(If ρ is a number, we mean $\rho_k = \rho$ for all k.) In two variables, a polydisc is sometimes called a *bidisc*. In particular, the *unit polydisc* is given by is

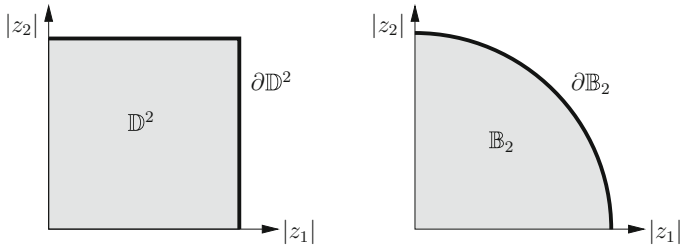
$$\mathbb{D}^n = \mathbb{D} \times \mathbb{D} \times \cdots \times \mathbb{D} = \Delta_1(0). \tag{1.1.3}$$

We will use the following notation for the Euclidean inner product on $\mathbb{C}^n$ and the standard Euclidean norm on $\mathbb{C}^n$:

$$\langle z, w \rangle = z \cdot \bar{w}\,, \qquad \|z\| = \sqrt{\langle z, z \rangle}. \tag{1.1.4}$$

Then we define $B_\rho(a)$ to be the ball in the metric $\|\cdot\|$. For example, $\mathbb{B}_n = B_1(0)$ is the *unit ball*.

Example 1.1.1 In more than one complex dimension, it is difficult to draw pictures because we lack dimensions on our paper. We instead draw pictures by plotting against the modulus of the variables. For example, the unit polydisc $\mathbb{D}^2$ and unit ball $\mathbb{B}_2$ in $n = 2$ complex dimensions can be visualized as follows:



(1.1.5)

Not every domain (by domain we mean an open connected set) can be drawn like this. If it can, it is called a *Reinherdt domain*.

The function $f\colon U \subset \mathbb{C}^n \to \mathbb{C}$ is called *holomorphic* if f is complex differentiable in each variable separately, that is, if

$$z_\ell \mapsto f(z_1, \ldots, z_\ell, \ldots, z_n) \text{ is complex differentiable for every } \ell\,. \tag{1.1.6}$$

Let us write $\mathcal{O}(U)$ to denote the set of holomorphic functions on U. Here, the letter $\mathcal{O}$ is used to recognize the fundamental contribution to several complex variables by Kiyoshi Oka.

From now on we will benefit from using the language of differential forms. Exterior derivative leads to 1-forms

$$\mathrm{d}z_\ell = \mathrm{d}x_\ell + i\,\mathrm{d}y_\ell, \qquad \mathrm{d}\bar{z}_\ell = \mathrm{d}x_\ell - i\,\mathrm{d}y_\ell. \tag{1.1.7}$$

As in one variable, we define the *Wirtinger operators*:

$$\frac{\partial}{\partial z_\ell} \overset{\text{def}}{=} \frac{1}{2}\left(\frac{\partial}{\partial x_\ell} - i\frac{\partial}{\partial y_\ell}\right), \qquad \frac{\partial}{\partial \bar{z}_\ell} \overset{\text{def}}{=} \frac{1}{2}\left(\frac{\partial}{\partial x_\ell} + i\frac{\partial}{\partial y_\ell}\right). \tag{1.1.8}$$

These are determined by being the dual bases of $\mathrm{d}z$ and $\mathrm{d}\bar{z}$

$$\mathrm{d}z_k\left(\frac{\partial}{\partial z_\ell}\right) = \delta_\ell^k, \quad \mathrm{d}z_k\left(\frac{\partial}{\partial \bar{z}_\ell}\right) = 0, \quad \mathrm{d}\bar{z}_k\left(\frac{\partial}{\partial z_\ell}\right) = 0, \quad \mathrm{d}\bar{z}_k\left(\frac{\partial}{\partial \bar{z}_\ell}\right) = \delta_\ell^k\,. \tag{1.1.9}$$

Alternatively, we might have said that f is holomorphic if it satisfies the *Cauchy–Riemann (CR) equations*:

$$\frac{\partial f}{\partial \bar{z}_\ell} = 0 \quad \text{for } \ell = 1, 2, \ldots, n\,. \tag{1.1.10}$$

If f is holomorphic, then its derivatives can be obtained by taking limits

$$\frac{\partial f}{\partial z_k}(z) = \lim_{\xi \in \mathbb{C} \to 0} \frac{f(z_1, \ldots, z_k + \xi, \ldots, z_n) - f(z)}{\xi}\,. \tag{1.1.11}$$

From now on, we will write a smooth function $f\colon U \subset \mathbb{C}^n \to \mathbb{C}$ simply as $f(z, \bar{z})$.

Example 1.1.2 If f is a polynomial (in x and y), write

$$x = \frac{z+\bar{z}}{2}, \qquad y = \frac{z-\bar{z}}{2i}, \tag{1.1.12}$$

and it really does become a polynomial in z and $\bar{z}$. E.g.,

$$2x_1 + 2y_1 + 4y_2^2 = (1-i)z_1 + (1+i)\bar{z}_1 - z_2^2 + 2z_2\bar{z}_2 - \bar{z}_2^2. \tag{1.1.13}$$

f is holomorphic if it does not depend on $\bar{z}$.

The derivatives satisfy the chain rule. Suppose $f\colon U \subset \mathbb{C}^n \to V \subset \mathbb{C}^m$, $g\colon V \to \mathbb{C}$ and that the variables are $z \in \mathbb{C}^n$ and $w \in \mathbb{C}^m$. Then

$$\frac{\partial}{\partial z_\ell}[g \circ f] = \sum_{k=1}^{m} \left(\frac{\partial g}{\partial w_k}\frac{\partial f_k}{\partial z_\ell} + \cancel{\frac{\partial g}{\partial \bar{w}_k}\frac{\partial \bar{f}_k}{\partial z_\ell}} \right), \tag{1.1.14}$$

$$\frac{\partial}{\partial \bar{z}_\ell}[g \circ f] = \sum_{k=1}^{m} \left(\cancel{\frac{\partial g}{\partial w_k}\frac{\partial f_k}{\partial \bar{z}_\ell}} + \cancel{\frac{\partial g}{\partial \bar{w}_k}\frac{\partial \bar{f}_k}{\partial \bar{z}_\ell}} \right) = 0, \tag{1.1.15}$$

provided that g and f are holomorphic.

Theorem 1.1.1 (Cauchy Integral Formula) *Let $\Delta \subset \mathbb{C}^n$ be a polydisc. Suppose $f\colon \overline{\Delta} \to \mathbb{C}$ is a continuous function, holomorphic in Δ, and that $\Gamma = \partial\Delta_1 \times \cdots \times \partial\Delta_n$ is oriented appropriately (each $\partial\Delta_k$ oriented positively). Then for $z \in \Delta$*

$$f(z) = \frac{1}{(2\pi i)^n} \int_\Gamma \frac{f(\zeta_1, \zeta_2, \ldots, \zeta_n)}{(\zeta_1 - z_1)(\zeta_2 - z_2)\cdots(\zeta_n - z_n)}\, d\zeta_1 \wedge d\zeta_2 \wedge \cdots \wedge d\zeta_n. \tag{1.1.16}$$

Here, we stated a general result where f is only continuous on $\bar{\Delta}$ and holomorphic on Δ. We are going to cheat and use the short-hand notation:

$$\frac{1}{\zeta - z} \stackrel{\text{def}}{=} \frac{1}{(\zeta_1 - z_1)(\zeta_2 - z_2)\cdots(\zeta_n - z_n)}, \tag{1.1.17}$$

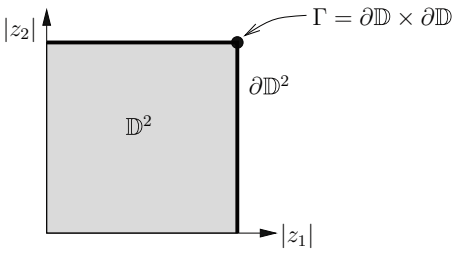
together with

$$\mathrm{d}\zeta \stackrel{\text{def}}{=} \mathrm{d}\zeta_1 \wedge \mathrm{d}\zeta_2 \wedge \cdots \wedge \mathrm{d}\zeta_n. \tag{1.1.18}$$

This allows us to write the Cauchy integral formula in a concise way:

$$f(z) = \frac{1}{(2\pi i)^n} \int_\Gamma \frac{f(\zeta)}{\zeta - z}\, \mathrm{d}\zeta. \tag{1.1.19}$$

The Cauchy integral formula shows an important and subtle point about the holomorphic functions in several variables: The function $f(z)$ for $z \in \Delta$ is determined in terms of its values on the set Γ, which is much smaller than the boundary of the polydisc $\partial\Delta$. In fact, Γ (a torus) is of real dimension n, while the boundary has real dimension $2n-1$. This is the first big difference compared to 1D. We call Γ the *distinguished boundary*. Let us draw the unit bidisc:



(1.1.20)

In this case, the set Γ is a 2-dimensional torus, like the surface of a donut. On the other hand, the set $\partial\mathbb{D}^2 = (\partial\mathbb{D} \times \overline{\mathbb{D}}) \cup (\overline{\mathbb{D}} \times \partial\mathbb{D})$ is the union of two filled donuts, or more precisely, it is both the inside and the outside of the donut put together, and these two things meet on the surface of the donut. So, the set Γ is quite small in comparison to the entire bounary $\partial\mathbb{D}^2$.

1.1.2 Power Series Representation

Writing expressions out in all the components can be a pain. For $\alpha \in \mathbb{N}_0^n$ (where $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$), we will cheat some more and use the multi-index notation to deal with the more complicated formulas:

$$z^\alpha \stackrel{\text{def}}{=} z_1^{\alpha_1} z_2^{\alpha_2} \cdots z_n^{\alpha_n}, \qquad \frac{\partial^{|\alpha|}}{\partial z^\alpha} \stackrel{\text{def}}{=} \frac{\partial^{\alpha_1}}{\partial z_1^{\alpha_1}} \frac{\partial^{\alpha_2}}{\partial z_2^{\alpha_2}} \cdots \frac{\partial^{\alpha_n}}{\partial z_n^{\alpha_n}}, \tag{1.1.21}$$

$$\alpha! \stackrel{\text{def}}{=} \alpha_1!\alpha_2! \cdots \alpha_n!, \qquad \alpha + 1 \stackrel{\text{def}}{=} (\alpha_1 + 1, \alpha_2 + 1, \cdots \alpha_n + 1). \tag{1.1.22}$$

Let Δ be a polydisc with distinguished boundary Γ, centered at a, of polyradius ρ. Suppose f is continuous on $\overline{\Delta}$, holomorphic on Δ. We will now differentiate under the integral in the Cauchy integral formula. This implies that f is infinitely $\mathbb{C}$-differentiable and

$$\frac{\partial^{|\alpha|} f}{\partial z^\alpha}(z) = \frac{1}{(2\pi i)^n} \int_\Gamma \frac{\alpha! f(\zeta)}{(\zeta - z)^{\alpha+1}} \, \mathrm{d}\zeta. \tag{1.1.23}$$

From this, we get *the Cauchy estimates*, which are bounds on the growth of derivatives of f:

$$\left|\frac{\partial^{|\alpha|} f}{\partial z^\alpha}(a)\right| \leqslant \frac{\alpha! \, \|f\|_\Gamma}{\rho^\alpha} = \frac{\alpha! \, \sup_{z\in\Gamma} |f(z)|}{\rho^\alpha}. \tag{1.1.24}$$

In particular, the coefficients of the power series depend only on the derivatives of f at a and not the specific polydisc Δ used above.

Corollary 1.1.1 *The "correct" topology on $\mathcal{O}(U)$ is uniform convergence on compacts (normal convergence). If $f_n \in \mathcal{O}(U)$ and $f_n \to f$ uniformly on compacts, then $f \in \mathcal{O}(U)$ and $f_n^{(\ell)} \to f^{(\ell)}$ uniformly on compacts.*

As in one variable, we can introduce the geometric series in several variables. For $z \in \mathbb{D}^n$ (unit polydisc):

$$\frac{1}{1-z} = \frac{1}{(1-z_1)(1-z_2)\cdots(1-z_n)} = \left(\sum_{k=0}^{\infty} {z_1}^k\right)\left(\sum_{k=0}^{\infty} {z_2}^k\right)\cdots\left(\sum_{k=0}^{\infty} {z_n}^k\right) \tag{1.1.25}$$

$$= \sum_{\alpha_1=0}^{\infty}\sum_{\alpha_2=0}^{\infty}\cdots\sum_{\alpha_n=0}^{\infty}\left({z_1}^{\alpha_1}{z_n}^{\alpha_2}\cdots{z_n}^{\alpha_n}\right) = \sum_{\alpha} z^\alpha. \tag{1.1.26}$$

Power series $\sum_\alpha c_\alpha (z-a)^\alpha$ converges absolutely uniformly on compact subsets of its *domain of convergence* (interior of the set where it converges).

Example 1.1.3 In $\mathbb{C}^2$, the power series

$$\sum_{k=0}^{\infty} z_1 z_2^k, \tag{1.1.27}$$

converges absolutely on the set

$$\left\{z \in \mathbb{C}^2 : |z_2| < 1\right\} \cup \left\{z \in \mathbb{C}^2 : z_1 = 0\right\}, \tag{1.1.28}$$

and nowhere else. This set is not a polydisc. It is neither an open set nor a closed set. Its closure is not the closure of the domain of convergence, which is the set $\left\{z \in \mathbb{C}^2 : |z_2| < 1\right\}$.

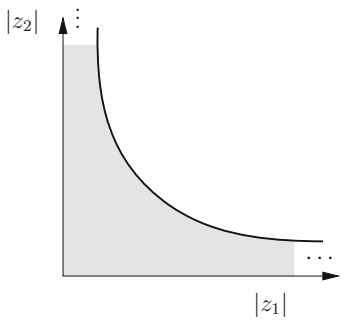
Example 1.1.4 The power series

$$\sum_{k=0}^{\infty} z_1^k z_2^k, \tag{1.1.29}$$

converges absolutely exactly on the set

$$\left\{z \in \mathbb{C}^2 : |z_1 z_2| < 1\right\}. \tag{1.1.30}$$

The picture of this domain is more complicated than that of a polydisc:



(1.1.31)

Let $\Delta = \Delta_\rho(a) \subset \mathbb{C}^n$ be a polydisc and f is holomorphic in a neighborhood of $\overline{\Delta}$, let Γ be the distinguished boundary of Δ. In the Cauchy integral formula,

$$f(z) = \frac{1}{(2\pi i)^n} \int_\Gamma \frac{f(\zeta)}{\zeta - z}\, \mathrm{d}\zeta\,, \tag{1.1.32}$$

let us expand the Cauchy kernel as

$$\frac{1}{\zeta - z} = \frac{1}{\zeta - a}\left(\frac{1}{1 - \frac{z-a}{\zeta - a}}\right) = \frac{1}{\zeta - a}\sum_\alpha \left(\frac{z-a}{\zeta - a}\right)^\alpha. \tag{1.1.33}$$

Here, we make sure to interpret the formulas correctly as $\frac{1}{\zeta - z} = \frac{1}{(\zeta_1 - z_1)\cdots(\zeta_n - z_n)}$ and so on. The multivariable geometric series is a product of geometric series in one variable, and geometric series in one variable are uniformly absolutely convergent on compact subsets of the unit disc. This allows us to prove the following theorem.

Theorem 1.1.2 *For $z \in \Delta$,*

$$f(z) = \sum_\alpha c_\alpha (z-a)^\alpha, \quad \text{where} \quad c_\alpha = \frac{1}{\alpha!}\frac{\partial^{|\alpha|} f}{\partial z^\alpha}(a) = \frac{1}{(2\pi i)^n}\int_\Gamma \frac{f(\zeta)}{(\zeta - a)^{\alpha+1}}\, d\zeta. \tag{1.1.34}$$

Conversely, if f is defined by a power series, then it is holomorphic.

The proof of the first statement is a simple computation and application of the Fubini theorem or uniform convergence just as in one variable. The converse statement

can be proven in different ways. For example, it follows by applying the Cauchy–Riemann equations to the series termwise.

This is in fact the first place where the theory of several complex variables becomes annoying.

Theorem 1.1.3 (Identity) *Let $U \subset \mathbb{C}^n$ be a domain (connected open set) and let $f : U \to \mathbb{C}$ be holomorphic. If $f|_N \equiv 0$ for a nonempty open subset $N \subset U$, then $f \equiv 0$.*

Here we encounter a difference from the 1D cases: The zero set of a holomorphic function in 2 or more complex variables is always large (it always has limit points). The above theorem is often used to show that if two holomorphic functions f and g are equal on a small open set, then $f \equiv g$.

Theorem 1.1.4 (Maximum Principle) *Let $U \subset \mathbb{C}^n$ be a domain. Let $f : U \to \mathbb{C}$ be holomorphic and suppose that $|f(z)|$ attains a local maximum at some $a \in U$. Then $f \equiv f(a)$.*

Here, the argument goes back to 1D. The proof involves using the maximum principle on every 1D subspace.

1.1.3 Inequivalence of Ball and Polydisc

We say that $f : U \to V$ is a *biholomorphism* (and U and V are *biholomorphic*) if f is bijective, holomorphic, and f^{-1} is holomorphic. One of the main questions in complex analysis is to classify domains up to biholomorphic transformations.

In one variable, there is the rather striking theorem due to Riemann: If $U \subset \mathbb{C}$ is a nonempty simply connected domain such that $U \neq \mathbb{C}$, then U is biholomorphic to $\mathbb{D}$. So essentially, in 1D a topological property on U is enough to classify a whole class of domains. It is one of the reasons why studying the disc is so important in one variable and why many theorems are stated for the disc only. There is no Riemann Mapping Theorem in several dimensions! In fact, it is not difficult to find 2D domains that are not biholomorphic.

Theorem 1.1.5 (Poincaré, 1907) *$\mathbb{B}_2$ and $\mathbb{D}^2$ are not biholomorphic.*

The first complete proof was given by Henri Cartan in 1931, though popularly the theorem is attributed to Poincaré. Note that both domains are simply connected (have no holes) and they are the two most obvious generalizations of the disc to two variables. They are homeomorphic, that is, topology does not see any difference.

We need to introduce some constructions before attempting a proof of the above theorem. A nonconstant holomorphic mapping $\varphi : \mathbb{D} \to \mathbb{C}^n$ is called an analytic disc. It plays an important role in the geometry of several complex variables. Essentially, it allows us to apply one-variable results in several variables.

It is especially important in understanding the boundary behavior of holomorphic functions and features prominently in complex geometry. Often, we call the image $\Delta = \varphi(\mathbb{D})$ the analytic disc rather than the mapping.

Proposition 1.1.1 *The unit sphere $S^{2n-1} = \partial\mathbb{B}_n \subset \mathbb{C}^n$ contains no analytic discs.*

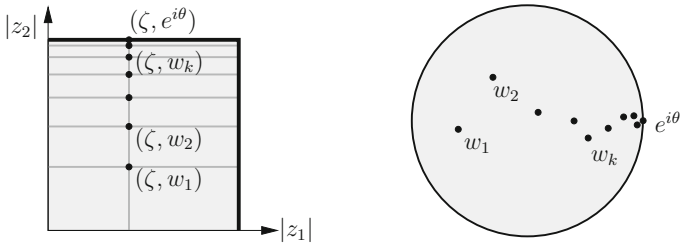
Proof *Suppose there is a holomorphic function $g\colon \mathbb{D} \to S^{2n-1} \subset \mathbb{C}^n$:*

$$|g_1(z)|^2 + |g_2(z)|^2 + \cdots + |g_n(z)|^2 = 1\,, \tag{1.1.35}$$

for all $z \in \mathbb{D}$. Without loss of generality (after composing with a unitary matrix) suppose that $g(0) = (1, 0, \cdots, 0)$. Notice that $g_1(0) = 1$ and $|g_1(z)| \leqslant 1$. The maximum principle says that g_1 attains its maximum for all $z \in \mathbb{D}$ which implies that $g_j(z) = 0$ for all $j = 2, \ldots, n$ on all $z \in \mathbb{D}$. Hence g has to be a constant function and not an analytic disc. □

The fact that the sphere contains no analytic discs is the most important geometric distinction between the boundary of the polydisc and the sphere.

Let us now sketch the proof of the Poincaré theorem using pictures. We will use proof by contradiction. To this end, suppose $f : \mathbb{D}^2 \to \mathbb{B}_2$ is a biholomorphism. Let us pick a disc for fixed $z_1 = \zeta$ and a sequence $w_k \to \mathrm{e}^{i\theta}$. It looks like this:



(1.1.36)

The idea is to show that (after passing to a subsequence via Montel) $\zeta \mapsto f(\zeta, w_k)$ converges to a holomorphic map to the sphere. Therefore, it has to be a constant. More precisely, the derivative of $\zeta \mapsto f(\zeta, w_k)$ goes to zero for every $\mathrm{e}^{i\theta}$ and every $\{w_k\}$. This implies that $\frac{\partial f}{\partial z_1} \equiv 0$ and by symmetry also $\frac{\partial f}{\partial z_2} \equiv 0$. Therefore f has to be a constant and we run into a contradiction.

The proof says that the reason why there is not even a proper mapping is the fact that the boundary of the polydisc contains analytic discs, whereas the sphere does not. The proof extends easily to higher dimensions, as well. The key takeaway is that in several complex variables, the geometry of the boundary makes a difference if one wants to determine if two domains are equivalent. The domain topology is not enough.

1.1.4 Cartan's Uniqueness Theorem

Where did Schwarz's lemma go? The following theorem is its analogue for several variables.

Theorem 1.1.6 (Cartan) *Suppose $U \subset \mathbb{C}^n$ is a **bounded** domain, $a \in U$, $f : U \to U$ is a holomorphic mapping, $f(a) = a$, and $Df(a)$ is the identity. Then $f(z) = z$ for all $z \in U$.*

Here, we bolded the word "bounded" which is crucial. Counterexamples can be found if U is unbounded.

The argument is to use Cauchy estimates on the first nonzero nonlinear term of the series of $f^\ell = f \circ f \circ \cdots \circ f$. The result can be used to compute automorphism groups just as in 1D. Every automorphism of $\mathbb{D}^n$ is of the form

$$z \mapsto P\left(e^{i\theta_1}\frac{a_1 - z_1}{1 - \bar{a}_1 z_1}, e^{i\theta_2}\frac{a_2 - z_2}{1 - \bar{a}_2 z_2}, \ldots, e^{i\theta_n}\frac{a_n - z_n}{1 - \bar{a}_n z_n}\right), \tag{1.1.37}$$

where $\theta \in \mathbb{R}^n$, $a \in \mathbb{D}^n$, and P is a permutation matrix. On the other hand, every automorphism of $\mathbb{B}_n$ is of the form

$$z \mapsto U\frac{a - P_a z - s_a(I - P_a)z}{1 - \langle z, a\rangle}, \tag{1.1.38}$$

where $a \in \mathbb{B}_n$, U is a unitary matrix, $s_a = \sqrt{1 - \|a\|^2}$, and $P_a z = \frac{\langle z,a\rangle}{\langle a,a\rangle}a$.

1.1.5 Riemann Extension Theorem, Zero Sets, and Injective Maps

In 1D, if a function is holomorphic in $U \setminus \{p\}$ and locally bounded in U (for every $q \in U$ there is a neighborhood W of q such that f is bounded on $W \cap (U \setminus \{p\})$), then the function extends holomorphically to U. In several variables, the same theorem holds, and the analogue of a single point is the zero set of a holomorphic function.

Theorem 1.1.7 (Riemann Extension Theorem) *Let $U \subset \mathbb{C}^n$ be a domain, $g \in \mathcal{O}(U)$, $g \not\equiv 0$, and $N = g^{-1}(0)$. If $f \in \mathcal{O}(U \setminus N)$ is locally bounded in U, then there exists a unique $F \in \mathcal{O}(U)$ such that $F|_{U\setminus N} = f$.*

Its proof involves cutting N "transversely" by complex lines, applying the 1D Riemann mapping theorem and using the Cauchy formula as glue.

The set of zeros of a holomorphic function has a nice structure at most points, as codified in the following theorem.

Theorem 1.1.8 *Let $U \subset \mathbb{C}^n$ be a domain, $f \in \mathcal{O}(U)$, $f \not\equiv 0$, and $N = f^{-1}(0)$. Then there exists an open and dense $N_{reg} \subset N$ such that at each $p \in N_{reg}$, after possibly reordering variables, N can be locally written as*

$$z_n = g(z_1, \ldots, z_{n-1}), \tag{1.1.39}$$

for a holomorphic function g.

The proof is to consider all possible derivatives of f. One of them must be nonzero somewhere on N (by analyticity). Then one applies the implicit function theorem.

For holomorphic $f : U \subset \mathbb{C}^n \to \mathbb{C}^n$, let us write the holomorphic Jacobian as

$$Df = \left[\frac{\partial f_k}{\partial z_\ell}\right]_{k\ell}. \tag{1.1.40}$$

Note that $|\det Df|^2 = \det D_{\mathbb{R}} f$, where $D_{\mathbb{R}} f$ is the real Jacobian matrix.

Theorem 1.1.9 *Suppose $U \subset \mathbb{C}^n$ is open and $f : U \to \mathbb{C}^n$ is holomorphic and one-to-one. Then* $\det Df$ *is never zero on U.*

In 1D, every holomorphic function f can, in the proper local holomorphic coordinates, be written as z^d for $d = 0, 1, 2, \ldots$, up to a constant. Such a simple result does not hold in several complex variables in general, but if the mapping is locally one-to-one, the present theorem says that such a mapping can be locally written as the identity. The proof of this theorem essentially reduces to the 1D statement, but not trivially.

Therefore, if a holomorphic map $f : U \to V$ is bijective for two open sets $U, V \subset \mathbb{C}^n$, then f is biholomorphic.

Example 1.1.5 The theorem does not hold in different dimensions. $f : \mathbb{C} \to \mathbb{C}^2$ given by $z \mapsto (z^2, z^3)$ is one-to-one and onto the cusp $(z_1^3 - z_2^2 = 0)$, but $f'(0) = 0$.

1.2 Convexity and Pseudoconvexity

Sean Curry

1.2.1 Pseudoconvexity

The motivating question of this lecture is: Which domains can be domains of maximal analytic continuation of a holomorphic function?

A central concept which will be introduced is *pseudoconvexity*. It will answer the above question. In fact, we will distinguish between three different notions:

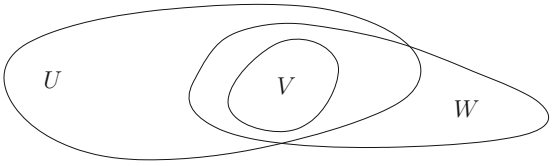
- Convexity (linear convexity in the usual sense),
- $\mathbb{C}$-linear convexity,
- Pseudoconvexity.[1]

For example, a bean shape is not convex but it may be pseudoconvex. We will elaborate on these notions in what follows.

In this talk, we will focus on domains rather than functions. It turns out that not every domain in $\mathbb{C}^n$ is a natural domain for holomorphic functions.[2]

Definition 1.2.1 Let $U \subseteq \mathbb{C}^n$ be a domain (a connected open set). The set U is called a *domain of holomorphy* if there do not exist nonempty open sets V and W of $\mathbb{C}^n$ with $V \subseteq U \cap W$, $W \not\subseteq U$, and W connected, such that for every function $f \in \mathcal{O}(U)$ there exists an $F \in \mathcal{O}(W)$ with $F|_V = f|_V$.

The idea here is that if a domain U is not a domain of holomorphy and both V, W exist as in the definition above, then f "extends across the boundary" somewhere. This is illustrated with the following picture, where V acts as "glue" between the other two domains:

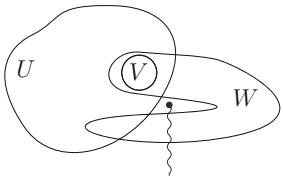


(1.2.1)

[1] In a tangential direction, Sean started by mentioning some places where the topics of Cauchy-Riemann geometry and (strong) pseudoconvexity arise in connection with physics ("tasty bits" that unfortunately go beyond the scope of these introductory lectures). The work of Fefferman [2–4] in the 1970s on strongly pseudoconvex domains inspired the work Fefferman and Graham [5] (cf. [6]) on conformal invariants, where they developed the boundary expansions for the asymptotically AdS metrics that are used in the AdS/CFT correspondence; Fefferman's result [2] also showed that the biholomorphically invariant geometry of a bounded strongly pseudoconvex domain in $\mathbb{C}^n$ is equivalent to the Cauchy-Riemann geometry of the boundary, a prototypical example of a "bulk-boundary correspondence" with strong connections to AdS/CFT. Cauchy-Riemann geometry also arises in various ways in twistor theory and in the antecedent work of Robinson, Trautman and others, where Cauchy-Riemann structures arise from shear-free null geodesic congruences in spacetimes (which are "twisting" if the Cauchy-Riemann structure is strongly pseudoconvex). In this setting a connection was found between Maxwell's equations and the tangential Cauchy-Riemann equations and then between Einstein's equations and the tangential Cauchy-Riemann equations. This played an important role in the discovery and study of algebraically special spacetimes such as the Kerr black hole solution [7–10].

[2] Sean is using the notation $\subseteq$ to mean the same as Jiří's $\subset$. Both symbols denote a subset that does not need to be proper.

Note that we also can have non-trivial monodromy as follows:



(1.2.2)

Here in principle the domain W can wrap around a branch cut, in which case the function "locally extends" around it. A function F defined on W is said to extend the function f on U if $F|_V = f|_V$ (for some open set $V \subseteq U \cap W$); the functions F and f may differ on other parts of $U \cap W$, i.e. the extension may give rise to a multivalued function. For example, the principal branch of the logarithm function on the cut plane is extendable across the branch cut away from the origin.

Remark 1.2.1 One can show (see, e.g., [11, Thm. 2.5.5]) that if Ω is a domain of holomorphy, then there exists a function $f \in \mathcal{O}(\Omega)$ that cannot be analytically continued past Ω. However, any individual $f \in \mathcal{O}(\Omega)$ might admit an analytic continuation outside of Ω.

Example 1.2.1 The unit disc $\mathbb{D}$ is a domain of holomorphy since we can always construct functions with singularities on any point of $\partial\mathbb{D}$. For example, if $\{1\} \in W$, then $\frac{1}{1-z}$ does not extend beyond $z = 1$. Similarly, if $\{e^{i\theta}\} \in W$, then $\frac{1}{e^{i\theta}-z}$ does not extend beyond $z = e^{i\theta}$.

Example 1.2.2 Alternatively, one can ask what functions have $\mathbb{D}$ as the maximal domain? We can construct such functions, with a dense set of singularities on the boundary $\partial\mathbb{D}$, by defining

$$f(z) = \sum_{k=0}^{\infty} z^{k!} \qquad \text{or} \qquad f(z) = \sum_{k=0}^{\infty} z^{a^k}\,, \tag{1.2.3}$$

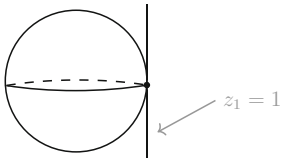
with $a \in \mathbb{Z}$, $a \geqslant 2$. Consider the first one. By plugging in $z = e^{2\pi i \frac{p}{q}}$ for any fraction $\frac{p}{q}$ we get that for any $k \geqslant q$ in the sum, $\left(e^{2\pi i \frac{p}{q}}\right)^{k!} = 1$, so the series defining $f(z)$ diverges at any $z = e^{2\pi i \frac{p}{q}}$ and from the sparsity of the coefficients we can conclude that $f(z)$ tends to infinity as z tends to any such boundary point. These points form a dense set of singularities on $\partial\mathbb{D}$, so that $f(z)$ cannot be analytically continued across any boundary point. Functions with such an obstruction to analytic continuations are called *lacunary* functions.

Remark 1.2.2 The argument from Example 1.2.1 showing that $\mathbb{D}$ is a domain of holomorphy can be applied to any domain $U \subseteq \mathbb{C}$. Thus, any domain $U \subseteq \mathbb{C}$ is a domain of holomorphy, since we can place poles at any point on the boundary of U.

Now we want to understand how domains of holomorphy can look like in several complex variables. Let us first consider an example.

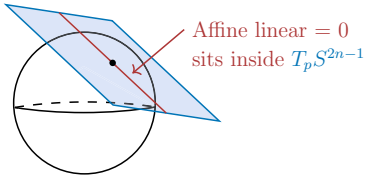
Example 1.2.3 The unit ball $\mathbb{B}^n \subseteq \mathbb{C}^n$ is a domain of holomorphy.

Proof If $(1, 0, \ldots, 0) \in W$, then the function $f(z) = \frac{1}{1-z_1}$ does not extend across $\mathbb{B}^n$.



(1.2.4)

By symmetry, other points can be handled by composing the function $f(z)$ with a rotation, giving a function of the form $\frac{1}{(\text{affine linear})}$ with a pole on a complex affine subspace of $\mathbb{C}^n$ that passes through the chosen point on $\partial\mathbb{B}^n$ (and is tangent to $\partial\mathbb{B}^n$ at that point).



(1.2.5)

Remark 1.2.3 The same construction clearly apples to any convex domain (recall that, for simplicity, we are restricting to domains with smooth boundary). In particular, any convex (or $\mathbb{C}$-linearly convex) domain $U \subseteq \mathbb{C}^n$ is a domain of holomorphy. (Here, $\mathbb{C}$-linearly convex means convex in complex tangential directions.)

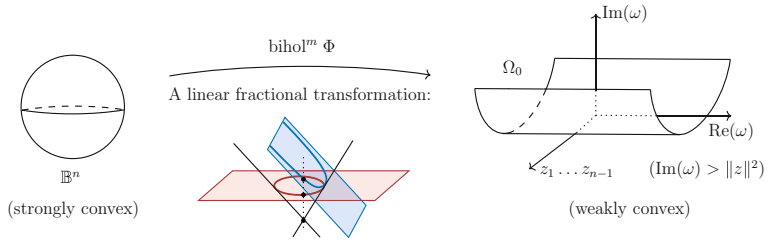
The key point is that, in order to have a domain of holomorphy in $\mathbb{C}^n$, we need room to "fit" the singularities which now lie on hypervarieties. Note that the ball $\mathbb{B}^n$ is convex and is in fact a domain of holomorphy. As noted in the preceding remark, this observation works more generally.

Theorem 1.2.1 *Any convex domain $U \subseteq \mathbb{C}^n$ is a domain of holomorphy.*

The question now becomes whether convexity is actually needed.

1.2.2 Non-convex Domain of Holomorphy

Domains of holomorphy are often not geometrically convex, so classical convexity is not the correct notion but it is in the right direction. Before giving an example of a non-convex domain of holomorphy, we show that by a biholomorphic change of coordinates the unit ball in $\mathbb{C}^n$ (which is strongly convex) can be realized as an unbounded domain that is convex but not strongly convex.



(1.2.6)

In the figure above, we map a ball by a biholomorphic linear fractional transformation[3] to the "Siegel upper half-space" Ω_0; this map (or its inverse) is known as the *(generalized) Cayley transform* as it generalizes the well-known Cayley map from the disc to the upper half-plane in one complex variable. Explicitly, the biholomorphism $\Phi : \mathbb{C}^n \setminus \{z_n = -1\} \to \mathbb{C}^n$ is defined by

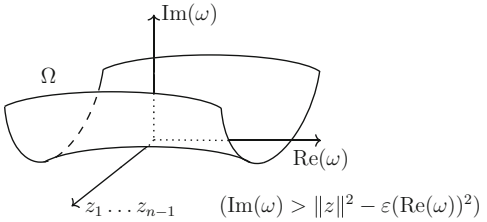
$$\Phi(z) = i\,\frac{e_n - z}{1 + z_n}$$

where $e_n = (0, \dots, 0, 1)$. (In the $n = 1$ case the map becomes $z \mapsto i\,\dfrac{1 - z}{1 + z}$, which is the inverse of the map $z \to \dfrac{i - z}{i + z}$ that takes the upper half plane to the disc.) Note that the domain on the left-hand side (the ball) is strongly convex, whereas the

[3] Geometrically, the map Φ can be seen as a change of affine chart for $\mathbb{CP}^n$. This is what is indicated in the figure under the arrow in (1.2.6) above. We think of the copy of $\mathbb{C}^n$ containing the ball on the left hand side of the biholomorphism above as the standard affine chart for complex projective space (the horizontal "plane" drawn in red in the figure below the arrow) so that $\mathbb{B}^n$ is identified with a subset of $\mathbb{CP}^n$ (a quadric, corresponding to the lines inside the cone in the picture below the arrow). The map Φ then simply amounts to viewing the ball $\mathbb{B}^n \subseteq \mathbb{CP}^n$ in a different affine chart for $\mathbb{CP}^n$ (corresponding to the "plane" drawn in blue that makes a 45° angle to the first, which we think of as the copy of $\mathbb{C}^n$ on the right hand side of the biholomorphism arrow that contains Ω_0). Note that Φ sends one point on $S^{2n-1} = \partial\mathbb{B}^n$ to infinity, namely the point $(0, \dots, 0, -1)$; this corresponds to the fact that the blue affine chart (the one drawn at 45°) cuts every line in the cone except the one corresponding to that point. The cone slicing picture very succinctly describes the generalized Cayley transform and shows the "parabolic" nature of the domain Ω_0 (slicing a cone at 45° gives a parabola). Unfortunately, however, we cannot draw in high enough dimensions to show how the flat direction in $\partial\Omega_0$ arises from slicing the "cone" in $\mathbb{C}^{n+1}$.

domain on the right-hand side (Ω_0) is only convex. If we perturb $\mathbb{B}^n$ slightly, it will remain strongly convex, but if we perturb Ω_0 slightly, it may become non-convex.

If we bend the "Siegel Upper Half Space" $\Omega_0 = \{\operatorname{Im} w > \|z\|^2\}$ as described above slightly to a domain Ω described by $\operatorname{Im} w > \varphi(z, \bar{z}, \operatorname{Re} w)$ where the function φ is defined by $\varphi(z, \bar{z}, \operatorname{Re} w) = \|z\|^2 - \epsilon(\operatorname{Re} w)^2$ for points near the origin (with $\epsilon > 0$ small) and is smoothly extended so that $\varphi(z, \bar{z}, \operatorname{Re} w) = \|z\|^2$ for points far from the origin, it remains a domain of holomorphy (assuming that the first and second derivatives of φ behave mildly on the interpolating region, which can easily be ensured). It stays convex in the directions $z_1, z_2, \ldots, z_n$, but not in the $\operatorname{Re} w$ direction:



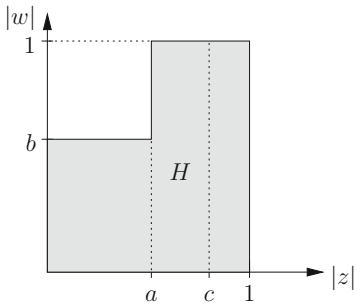
(1.2.7)

This is a non-convex example of a domain of holomorphy. Although it is not convex, one can still find complex hypervariety tangent to each point of $\partial\Omega$ that otherwise lies on the outside. For example, we can use the Cayley transformation to view this as a perturbation of $\mathbb{B}^n$ (which is convex) and then carry the singular functions over to handle the domain Ω.

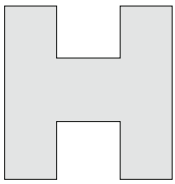
The above discussion means that what we really want is some kind of a notion of convexity "in the complex tangent directions." But only roughly: this would lead us to the notion of $\mathbb{C}$-linear convexity which is stronger than what we really want. We also want a notion of "convexity" that is invariant under (local) biholomorphic changes of coordinates. Before we come to the appropriate definition we will first consider some non-domains of holomorphy. A domain fails to be convex if there are two points in the domain such that the line segment between these two points contains points outside of the domain. We will see that a domain fails to be a domain of holomorphy if there is a complex disc (or more generally the image of a closed analytic disc $\varphi : \overline{\mathbb{D}} \to \mathbb{C}^n$) whose boundary lies inside the domain but whose interior contains points outside the domain, provided this disc can be obtained as a member of a continuous family of discs where one of the (closed) discs in the family is completely contained in the domain.

1.2.3 Non-domains of Holomorphy

Now that we have encountered a number of domains of holomorphy, it is time to ask when does a domain fail to be one? As a concrete example, consider the following *Hartogs figure* $H \subseteq \mathbb{C}^2$ (with $0 < a, b < 1$):



In diagrams, the Hartogs figure is often drawn as:



(1.2.8)

In drawing the diagram on the right, we have to imagine that H revolves around in (z, w) space with the radii given by the left picture. In particular, the "arms" of the "H" are connected to each other and contain circles that surround the white space between the arms. These circles bound a family of discs that cover all the points corresponding to the white space between the arms, and this is the key to the following theorem.

Theorem 1.2.2 (Hartogs, 1906) *Every $f \in \mathcal{O}(H)$ extends to $F \in \mathcal{O}(\mathbb{D} \times \mathbb{D})$.*

Proof *For every $f \in \mathcal{O}(H)$, we can use Cauchy's integral formula in z to extend it to $F \in \mathcal{O}(\mathbb{D} \times \mathbb{D})$ using the analytic dependence of boundary values on w.*

Hence, H is not a domain of holomorphy. For example, if we tried to put a pole somewhere in the middle, it would always poke out through the H. Hence, a pole contained in $\mathbb{D} \times \mathbb{D} \setminus H$ cannot exist. If we hypothetically had an isolated singularity, we could avoid H. Hence, a holomorphic function F cannot have isolated singularities.

There are higher-dimensional versions of the above figure with $w = (w_1, \ldots, w_k)$ and $z = (z_1, \ldots, z_n)$. Moreover, we can dilate and rotate this example using unitary transformations to "fit" it near the boundaries of other domains. For example, one can use it to show that $\mathbb{C}^2 \setminus \mathbb{R}^2$ is not a domain of holomorphy, but $\mathbb{C}^2 \setminus (\mathbb{C} \times \{0\})$ is. In another special case, we have the following result on removable singularities.

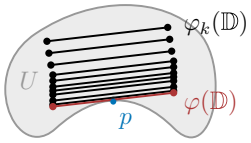
Corollary 1.2.1 *For $n > 1$, every $f \in \mathcal{O}(\mathbb{C}^n \setminus \{0\})$ extends to $F \in \mathcal{O}(\mathbb{C}^n)$.*

Audience Question 1.2.1 If we have U that is a domain of holomorphy but not convex, can we make a biholomorphic mapping to find a convex domain?
Answer: Not in general. It is a very hard problem. We'll see that domains of holomorphy are characterized by pseudoconvexity (meaning that the Levi form, which will be defined below, is nonnegative at every point). It is easy to show that a domain that is *strongly pseudoconvex* (meaning that the Levi form is everywhere positive definite) can be locally made strongly convex by a biholomorphic change of coordinates, but this can not always be done globally. On the other hand, even

locally not every pseudoconvex domain can be made convex by a biholomorphic change of coordinates, see [12, 13].

The key idea behind the Hartogs extension phenomenon described above is that one has a family of discs whose boundaries stay in the domain U but whose interiors sweep out the region to which one is trying to extend. The next theorem, the Kontinuitätssatz, is essentially saying that to establish the local holomorphic extension of functions $f \in \mathcal{O}(U)$ accross a boundary point $p \in \partial U$ it is enough to find a family of holomorphic discs that remain inside U (and whose boundaries stay "well inside") tending to a limiting disc that contains p in its interior.

Theorem 1.2.3 (Kontinuitätssatz, Version 1) *Suppose that $U \subseteq \mathbb{C}^n$ is open and there exists a sequence of closed analytic discs $\varphi_k : \overline{\mathbb{D}} \to \mathbb{C}^n$ converging (pointwise) to a closed analytic disc $\varphi : \overline{\mathbb{D}} \to \mathbb{C}^n$, such that $\varphi_k(\overline{\mathbb{D}}) \subseteq U$ for all k and $\varphi(\partial\mathbb{D}) \subseteq U$. Then there exists a polyradius $s > 0$ such that for every $f \in \mathcal{O}(U)$ and every $p \in \varphi(\mathbb{D})$, there is an $F \in \mathcal{O}(\Delta_s(p))$ with $F = f$ on some open subset of $U \cap \Delta_s(p)$.*



(1.2.9)

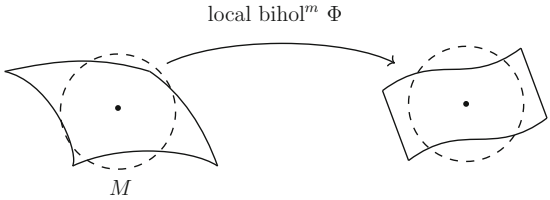
In the above figure the theorem gives us a fixed neighborhood $\Delta_s(p)$ of the point p such that every holomorphic function f in U extends to $\Delta_s(p)$.

The idea behind the proof is as follows. The boundary points of the discs stay bounded away from ∂U and the Cauchy estimates at/near the boundary points propagate to the interior of the discs by the maximum principle. Thus, uniform lower bounds for the "polyradii" of convergence are obtained, giving the result.

Now, we want to find out what is the right condition on ∂U (assuming that it is smooth) to avoid the presence of the above discs. This question can be studied with the Cauchy–Riemann (CR) structure of the boundary.

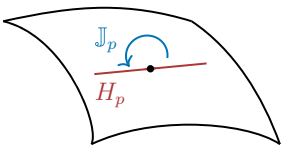
1.2.4 The CR Geometry of the Boundary

Consider a local biholomorphism Φ from M to some M'. Which part of the geometry is preserved?



(1.2.10)

The derivative $D\Phi$ is a $\mathbb{C}$-linear isomorphism at each point. It preserves the *maximal complex subspace* $H_p \subseteq T_pM$ and the *complex structure* $J_p : H_p \to H_p$, the restriction of the ambient complex structure $\mathbb{J}_p : T_p\mathbb{C}^n \to T_p\mathbb{C}^n$ to $H_p = T_pM \cap \mathbb{J}_p T_pM$. Since $\mathbb{J}_p^2 = -\mathrm{Id}$ on $T_p\mathbb{C}^n$, $J_p^2 = -\mathrm{Id}$ on H_p. Vectors not lying in H_p get rotated out of the tangent space T_pM when acted on by $\mathbb{J}_p$:



(1.2.11)

So the part of the geometry that is preserved is the *CR structure*: (M, H, J). Although J_p allows us to think of the real $2(n-1)$ dimensional vector space H_p as a complex vector space (defining the multiplication by i by $iX = J_pX$ for $X \in H_p$) it is customary and in the end much more convenient to think of H_p as being a real vector space. Since the operator $J_p : H_p \to H_p$ squares to minus the identity, it has eigenvalues $\pm i$; hence, if we complexify H_p to $\mathbb{C}H_p = \mathbb{C} \otimes_{\mathbb{R}} H_p$ we may decompose $\mathbb{C}H_p$ into i and $-i$ eigenspaces as $\mathbb{C}H_p = T_p^{(1,0)}M \oplus T_p^{(0,1)}M$. Here, $T_p^{(1,0)}M$ is the $(n-1)$-dimensional complex vector space $\{X - iJ_pX \mid X \in H_p\}$ and $T_p^{(0,1)}M$ is the conjugate vector space $\{X + iJ_pX \mid X \in H_p\}$. Equivalently, $T_p^{1,0}M = T_p^{(1,0)}\mathbb{C}^n \cap \mathbb{C}T_pM = \ker \partial r|_p$ where r is a local defining function for M and

$$T_p^{(1,0)}\mathbb{C}^n = \left\{ \frac{\partial}{\partial z_1}\bigg|_p, \ldots, \frac{\partial}{\partial z_n}\bigg|_p \right\}. \tag{1.2.12}$$

Knowing $T_p^{(1,0)}M$ is equivalent to knowing H_p and J_p so this gives an alternative way to define the CR structure. For convenience we will write $d = n-1$ for the complex dimension of $T_p^{(1,0)}M$, which is known as the *CR dimension* of M. The space $T_p^{(1,0)}M$ is known as the *holomorphic tangent space* to M at p. Typically $T_p^{(1,0)}M$ is described by giving an explicit frame, see, e.g., Examples 1.4.1 and 1.4.2 in Lecture 4.

The key point is this. If we pick a complementary direction T (a vector field tangent to M) to $H_p \subseteq T_pM$, then the direct sum decomposition

$$T_pM = H_p \oplus \mathbb{R}T_p\,, \tag{1.2.13}$$

complexifies to

$$\mathbb{C}T_pM = T_p^{(1,0)}M \oplus T_p^{(0,1)}M \oplus \mathbb{C}T_p\,. \tag{1.2.14}$$

The final decomposition can be written as $\mathcal{E}^{\alpha} \oplus \mathcal{E}^{\overline{\alpha}} \oplus \mathcal{E}^0$ in Penrose-style notation. If r is a local defining function for $M = \partial U$ (meaning that $r < 0$ in U while $r = 0$ on ∂U and $dr \neq 0$ on ∂U), we can write the corresponding decomposition of the

tangential part of its Hessian using Penrose abstract index notation as:

$$\operatorname{Hess}(r)\big|_{T_pM} = \begin{pmatrix} r_{\alpha\beta} & r_{\alpha\bar{\beta}} & r_{\alpha 0} \\ r_{\bar{\alpha}\beta} & r_{\bar{\alpha}\bar{\beta}} & r_{\bar{\alpha}0} \\ r_{0\beta} & r_{0\bar{\beta}} & r_{00} \end{pmatrix}. \tag{1.2.15}$$

Here, $r_{\alpha\overline{\beta}}$ and $r_{\overline{\alpha}\beta}$ are conjugates, as are $r_{\alpha\beta}$ and $r_{\overline{\alpha\beta}}$. In this notation, convexity means that this Hessian is positive (in the tangential directions):

$$\operatorname{Hess}(r)\big|_{T_pM} \geqslant 0 \quad \text{for all } p \in M\,. \tag{1.2.16}$$

On the other hand, pseudoconvexity means that only

$$r_{\alpha\overline{\beta}} \geqslant 0 \quad \text{for all } p \in M\,. \tag{1.2.17}$$

Remark 1.2.4 The Penrose-style notation is meant to be interpreted as expressing the abstract block decomposition of $\operatorname{Hess}(r)\big|_{T_pM}$ corresponding to the decomposition (1.2.14) of the (complexified) tangent space. But, of course, if one were to introduce a basis $(Z_1, \ldots, Z_d)$ for $T^{(1,0)}_p M$ and set $Z_{\bar{\alpha}} = \overline{Z_\alpha}$, then $(Z_1, \ldots, Z_d, Z_{\bar{1}}, \ldots, Z_{\bar{d}}, T)$ would be a basis of $\mathbb{C}T_pM$ that respects the decomposition (1.2.14) and (1.2.15) could then be interpreted as giving the $(2d+1)\times(2d+1)$ matrix for $\operatorname{Hess}(r)\big|_{T_pM}$; in this case we can interpret $r_{\alpha\overline{\beta}}$ as $\operatorname{Hess}(r)(Z_\alpha, Z_{\bar{\beta}})$ and condition (1.2.17) is the condition that the matrix $(r_{\alpha\overline{\beta}})$ is positive semidefinite.

Finally, let us wrap up this lecture with a precise definition.

Definition 1.2.2 (Levi Pseudoconvexity) Suppose $U \subseteq \mathbb{C}^n$ is an open set with smooth boundary, and r is a defining function for U at p with $r < 0$ in U. Then U is said to be *pseudoconvex* at $p \in M = \partial U$ if

$$\sum_{j,k=1}^{n} a_j \bar{a}_k \frac{\partial^2 r}{\partial z_j \partial \bar{z}_k}\bigg|_p \geqslant 0 \qquad \text{for all} \qquad X_p = \underbrace{\sum_{j=1}^{n} a_j \frac{\partial}{\partial z_j}\bigg|_p}_{X_p r = \sum_{j=1}^n a_j \frac{\partial r}{\partial z_j}\big|_p = 0} \in T^{(1,0)}_p M\,. \tag{1.2.18}$$

The hermitian form on $T^{(1,0)}_p M$ defined by the left hand side of the above display is called the *Levi form* of $M = \partial U$ at p. So, a domain is said to be pseudoconvex at $p \in \partial U$ if and only if the Levi form of ∂U is positive semidefinite at p.

This is the biholomorphically invariant part of the convexity condition. Given a point $p \in M$ (using $dr_p \neq 0$) one can find a local biholomorphic change

of coordinates so that in the new coordinates ζ we have $\frac{\partial^2 r}{\partial\zeta_j\partial\zeta_k}\Big|_p = 0$. On the other hand, the *complex Hessian* $\frac{\partial^2 r}{\partial z_j\partial\bar{z}_k}$ transforms as a tensor under change of coordinates: if $\zeta_m = f_m(z_1,\dots,z_n), m = 1,\dots,n$, is a local biholomorphic change of variables and $f = (f_1,\dots,f_m)$ then from the chain rule and the fact that $\frac{\partial f_m}{\partial\bar{\zeta}_j} = 0$ one can easily show that

$$\frac{\partial^2(r\circ f)}{\partial\zeta_j\partial\bar{\zeta}_k} = \sum_{\ell,m=1}^{n} \frac{\partial^2 r}{\partial z_\ell\partial\bar{z}_m}\frac{\partial f_\ell}{\partial\zeta_j}\frac{\partial\overline{f_m}}{\partial\bar{\zeta}_k}, \tag{1.2.19}$$

(the corresponding formula for $\frac{\partial^2(r\circ f)}{\partial\zeta_j\partial\zeta_k}$ involves the second derivatives of f and from the explicit expression one can see that at a point p where $dr \neq 0$ one can find a change of coordinates map f such that $\frac{\partial^2(r\circ f)}{\partial\zeta_j\partial\zeta_k}$ is zero at p).

What is the significance in restricting the complex Hessian appearing in (1.2.18) to holomorphic tangent vectors $X_p \in T^{(1,0)}_p M$ to M (rather than considering all $X_p \in T^{(1,0)}_p\mathbb{C}^n$)? The key point here is that $\sum_{j,k=1}^n a_j\bar{a}_k \frac{\partial^2 r}{\partial z_j\partial\bar{z}_k}\Big|_p$ only behaves nicely under a change in the defining function r when $X_p = \sum_{j=1}^n a_j \frac{\partial}{\partial z_j}\Big|_p \in T^{(1,0)}_p M$ (equivalently, when $X_p r = \sum_{j=1}^n a_j \frac{\partial r}{\partial z_j}\big|_p = 0$): If $\tilde{r} = \mathrm{e}^{\Upsilon} r$ is another defining function (where Υ is an arbitrary smooth real valued function) then, since $r(p) = 0$,

$$\sum_{j,k=1}^{n} a_j\bar{a}_k\frac{\partial^2\tilde{r}}{\partial z_j\partial\bar{z}_k}\Big|_p = \mathrm{e}^{\Upsilon}\sum_{j,k=1}^{n}\left(a_j\bar{a}_k\frac{\partial^2 r}{\partial z_j\partial\bar{z}_k} + a_j\frac{\partial\Upsilon}{\partial z_j}a_k\frac{\partial r}{\partial\bar{z}_k} + a_j\frac{\partial r}{\partial z_j}a_k\frac{\partial\Upsilon}{\partial\bar{z}_k}\right)\Big|_p. \tag{1.2.20}$$

When $X_p \in T^{(1,0)}_p M$ this simplifies to

$$\sum_{j,k=1}^{n} a_j\bar{a}_k\frac{\partial^2\tilde{r}}{\partial z_j\partial\bar{z}_k}\Big|_p = \mathrm{e}^{\Upsilon}\sum_{j,k=1}^{n} a_j\bar{a}_k\frac{\partial^2 r}{\partial z_j\partial\bar{z}_k}\Big|_p, \tag{1.2.21}$$

and hence the condition (1.2.18) in Definition 1.2.2 does not depend on the choice of defininig function r. Moreover, the Levi form of $M = \partial U$ at p is well defined (independent of the choice of the defining function) up to multiplication by a positive constant.

Remark 1.2.5 Some of the material covered at the start of Lecture 4 (Sect. 1.4) was meant to be covered before Jiří's lecture on CR functions (Sect. 1.3), but there was not enough time. The reader may find it helpful to look at these parts of Lecture 4 (Sects. 1.4.1 and 1.4.2) before reading Sect. 1.3.

1.3 CR Functions

Jiří Lebl

1.3.1 Real Analytic Functions and Complexification

Let us recall a simple result, which is a traditional way of interpreting complexification.

If $U \subset \mathbb{C}^n$ is a domain, $U \cap \mathbb{R}^n \neq \emptyset$, $f, g \in \mathcal{O}(U)$, and $f = g$ on $U \cap \mathbb{R}^n$, then $f \equiv g$. The result goes the other way as well: If $V \subset \mathbb{R}^n$, $f \colon V \to \mathbb{R}$ is real-analytic (that is, locally given by real power series), then there exists an open domain $U \subset \mathbb{C}^n$ such that $V \subset U$ and a unique holomorphic function $F \in \mathcal{O}(U)$ such that $F|_V = f$. The proof is essentially that given real power series $\sum_\alpha c_n (x-p)^n$, we can insert complex numbers and obtain $\sum_\alpha c_n (z-p)^n$.

There is usually a better way to complexify real-analytic functions in $\mathbb{C}^n$. Suppose $U \subset \mathbb{C}^n \cong \mathbb{R}^{2n}$ and $f \colon U \to \mathbb{C}$ is real-analytic. Write (at 0 for simplicity)

$$f(x,y) = \sum_{m=0}^{\infty} f_m(x,y) = \sum_{m=0}^{\infty} f_m\left(\frac{z+\bar{z}}{2}, \frac{z-\bar{z}}{2i}\right). \tag{1.3.1}$$

The polynomial f_m becomes a homogeneous polynomial of degree m in the variables z and $\bar{z}$, which means the series is a power series in z and $\bar{z}$. Hence, at any point, the function f equals

$$\sum_{\alpha,\beta} c_{\alpha,\beta} (z-a)^\alpha (\bar{z}-\bar{a})^\beta, \tag{1.3.2}$$

in multinomial notation. We will simply write it as $f(z,\bar{z})$. A holomorphic function is real-analytic, but not vice versa. A holomorphic function is a real-analytic function that does not depend on $\bar{z}$.

Before further discussion, we will need the following result. Let $U \subset \mathbb{C}^n \times \mathbb{C}^n$ be a domain and $f, g \in \mathcal{O}(U)$ such that $f = g$ on the *diagonal*

$$U \cap D = U \cap \left\{(z,\zeta) \in \mathbb{C}^n \times \mathbb{C}^n : \zeta = \bar{z}\right\}. \tag{1.3.3}$$

Then $f \equiv g$ on all of U. The result also goes the other way: If $f \colon V \subset D \to \mathbb{C}$ is real-analytic, then f extends to a neighborhood of V in $\mathbb{C}^{2n}$. We identify $\mathbb{C}^n$ and $D \subset \mathbb{C}^n \times \mathbb{C}^n$ with $\iota(z) = (z,\bar{z})$.

Example 1.3.1 In one dimension, the function

$$f(z,\bar{z}) = \frac{1}{1+|z|^2} = \frac{1}{1+z\bar{z}}, \tag{1.3.4}$$

is real-analytic in $\mathbb{C}$, but is not a restriction on the diagonal of a holomorphic function in all of $\mathbb{C}^2$. The problem is that the complexified function

$$f(z,\zeta) = \frac{1}{1+z\zeta}\,, \tag{1.3.5}$$

is holomorphic in $\mathbb{C}^2 \setminus \{z\zeta = -1\}$, i.e., it is undefined on the set where $z\zeta = -1$.

Example 1.3.2 If $u(z,\bar{z})$ is a (pluri)harmonic function, then $u(z,\bar{z}) = \operatorname{Re} f(z)$. How can we recover f from u in this case? First, notice that we have

$$u(z,\bar{z}) = \frac{f(z) + \bar{f}(\bar{z})}{2}\,. \tag{1.3.6}$$

Without loss of generality, we can set $f(0) = 0$. This implies

$$f(z) = 2u(z,0)\,. \tag{1.3.7}$$

The idea of treating $\bar{z}$ as a separate variable is very powerful, and as we have just seen, it is completely natural when speaking about real-analytic functions.[4] This is one of the reasons why real-analytic functions play a special role in several complex variables.

Remark 1.3.1 There is no good control over the neighborhood to which f extends. This is true even in 1D: Given any interval (a,b) and any neighborhood U of (a,b), there is a function $F \in \mathcal{O}(U)$ that does not extend beyond any boundary point of U. So $f = F|_{(a,b)}$ also cannot extend further. If we want any additional estimates to hold, we need to know something more about the function.

1.3.2 CR Functions

So far we have talked about the submanifold $\mathbb{R}^n \subset \mathbb{C}^n$. What can we say about more complicated submanifolds?

Suppose $M \subset \mathbb{C}^n$ is a hypersurface, then a function $f\colon M \to \mathbb{C}$ is a *CR function* if

$$X_p f = 0\,, \tag{1.3.8}$$

for all vectors $X_p \in T^{(0,1)}_p M$ for all points $p \in M$. Moreover, if $M \subset U \subset \mathbb{C}^n$ and $F \in \mathcal{O}(U)$, then $F|_M$ is a CR function.

[4] Jiří recalls a story of asking his advisor if there is a typo in the formula (1.3.7) because it looks too good to be true.

The question is whether the reverse statement holds. In fact, it is not always true: If M is real-analytic, then also $F|_M$ is real-analytic, so no smooth-only CR function f on M is such a restriction.

Theorem 1.3.1 (Severi) *If M and f are real-analytic and f CR, then f extends holomorphically to a neighborhood.*

Proof *The proof feels like cheating, so let us do it. Suppose without loss of generality that $0 \in M$ and M is real-analytic. Then, there is a holomorphic function $\Phi(z, \zeta, w)$ in a neighborhood of 0 in $\mathbb{C}^{n-1} \times \mathbb{C}^{n-1} \times \mathbb{C}$, such that M is*

$$\bar{w} = \Phi(z, \bar{z}, w), \tag{1.3.9}$$

and Φ, $\frac{\partial \Phi}{\partial z_k}$, $\frac{\partial \Phi}{\partial \zeta_k}$ vanish at 0 and $w = \bar{\Phi}\big(\zeta, z, \Phi(z, \zeta, w)\big)$. A basis for $T^{(0,1)}M$ is given by

$$\frac{\partial}{\partial \bar{z}_k} + \frac{\partial \Phi}{\partial \bar{z}_k}\frac{\partial}{\partial \bar{w}} \quad \left(= \frac{\partial}{\partial \bar{z}_k} + \frac{\partial \Phi}{\partial \zeta_k}\frac{\partial}{\partial \bar{w}} \right), \qquad k = 1, \ldots, n-1. \tag{1.3.10}$$

We therefore conclude that M is $\bar{w} = \Phi(z, \bar{z}, w)$, $T^{(0,1)}M$ is given by $\frac{\partial}{\partial \bar{z}_k} + \frac{\partial \Phi}{\partial \bar{z}_k}\frac{\partial}{\partial \bar{w}}$. Define the complexification $\mathcal{M} \subset \mathbb{C}^{2n}$ by $\omega = \Phi(z, \zeta, w)$. Moreover, let us complexify $f(z, w, \bar{z}, \bar{w})$ to $f(z, w, \zeta, \omega)$.

Now comes the trick: Let us define

$$F(z, w, \zeta) = f\big(z, w, \zeta, \Phi(z, \zeta, w)\big). \tag{1.3.11}$$

Because f is a CR function, it is killed by $\frac{\partial}{\partial \bar{z}_k} + \frac{\partial \Phi}{\partial \bar{z}_k}\frac{\partial}{\partial \bar{w}}$ on M. Hence

$$\frac{\partial F}{\partial \zeta_k} + \frac{\partial \Phi}{\partial \zeta_k}\frac{\partial F}{\partial \omega} = \frac{\partial F}{\partial \zeta_k} = 0. \tag{1.3.12}$$

This is true everywhere by complexification. So F is a function of z and w only. Therefore, F is a holomorphic function on $\mathbb{C}^n$ (on some neighborhood of M). □

Example 1.3.3 Consider $M \subset \mathbb{C}^2$ given by

$$\operatorname{Im} w = |z|^2 . \tag{1.3.13}$$

That is, $\frac{w-\bar{w}}{2i} = z\bar{z}$, or in other words, $\mathcal{M}$ is given by $\omega = -2iz\zeta + w$, and the CR vector field by $\frac{\partial}{\partial \bar{z}} - 2iz\frac{\partial}{\partial \bar{w}}$.

The most important place where we find CR functions that are not necessarily real-analytic is as boundary values of holomorphic functions.

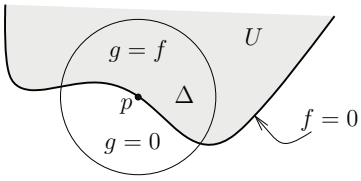
Proposition 1.3.1 *Suppose $U \subset \mathbb{C}^n$ is open with smooth boundary and $f : \overline{U} \to \mathbb{C}$ is smooth and holomorphic on U. Then $f|_{\partial U}$ is a smooth CR function.*

Proof *Each $X_p \in T_p^{(0,1)}\partial U$ is a limit of $T^{(0,1)}\mathbb{C}^n$ vectors from the inside.* □

The boundary values of a holomorphic function define the function uniquely. That is, if two holomorphic functions continuous up to the (smooth) boundary are equal on an open set of the boundary, then they are equal in the domain. This statement is made precise in the following proposition.

Proposition 1.3.2 *Suppose $U \subset \mathbb{C}^n$ is a domain with smooth boundary and $f : \overline{U} \to \mathbb{C}$ is smooth, holomorphic on U and $f|_{\partial U}$ is zero on a nonempty open subset. Then $f \equiv 0$.*

Proof Take $p \in \partial U$ such that $f = 0$ on a neighborhood of p in ∂U. Consider a small neighborhood Δ of p such that f is zero on $\partial U \cap \Delta$. Define $g : \Delta \to \mathbb{C}$ by setting $g(z) = f(z)$ if $z \in U$ and $g(z) = 0$ otherwise.



(1.3.14)

It is not hard to see that g is continuous, and it is clearly holomorphic where is not zero. Use Radó's theorem (see below) to extend g as 0 outside, then use the identity theorem.

Theorem 1.3.2 (Radó) *If $U \subset \mathbb{C}^n$ is open and $g : U \to \mathbb{C}$ continuous and holomorphic on*

$$U' = \{z \in U : g(z) \neq 0\}. \tag{1.3.15}$$

Then $g \in \mathcal{O}(U)$.

This is Jiří's favorite theorem.

1.3.3 Approximation of CR Functions

A problem we tackle next is trying to extend a smooth CR function from the boundary of a domain to a holomorphic function inside of it. This is essentially a PDE problem where the PDE are the Cauchy–Riemann equations, and the function on the boundary sets the boundary conditions. But notice that Cauchy–Riemann

equations are *overdetermined*, that is, there are too many equations. It means that not every boundary data gives a solution. The two propositions above say that, respectively, the boundary data being CR is a necessary condition for a solution (it is not sufficient in general) and that a solution is unique if it exists.

Let us give examples of functions that are not boundary values of holomorphic functions.

Example 1.3.4 Suppose $M = \mathbb{R} \subset \mathbb{C}$. Let us define the following function $f : M \to \mathbb{C}$:

$$f(x) = \begin{cases} e^{-x^{-2}} & \text{if } x \neq 0, \\ 0 & \text{if } x = 0. \end{cases} \tag{1.3.16}$$

Then f is CR (trivially), but is not a restriction nor a boundary value (from either side) of a holomorphic function continuous up to 0. We can come up with generalizations of this example to several variables by working on $M = \mathbb{R} \times \mathbb{C}$.

Example 1.3.5 Define the function $f \in \overline{\mathbb{B}}_2 \to \mathbb{C}$ by

$$f(z_1, z_2) = \begin{cases} e^{-1/\sqrt{z_1+1}} & \text{if } z_1 \neq -1, \\ 0 & \text{if } z_1 = -1. \end{cases} \tag{1.3.17}$$

Then f is smooth on $\overline{\mathbb{B}}_2$, holomorphic on $\mathbb{B}_2$, but near $(-1, 0)$ it is not a restriction of a holomorphic function (it is only a one sided extension).

A neat technique for extension is to approximate functions by polynomials. The following theorem holds in much more generality, but here we state its simplest version.

Theorem 1.3.3 (Baouendi–Trèves) *Suppose $M \subset \mathbb{C}^n$ is a smooth real hypersurface, $p \in M$. Then there exists a compact neighborhood $K \subset M$ of p, such that for every CR function $f : M \to \mathbb{C}$, there exists a sequence $\{p_\ell\}$ of polynomials in z such that*

$$p_\ell(z) \to f(z) \qquad \textit{uniformly in } K. \tag{1.3.18}$$

A key point is that K cannot be chosen arbitrarily, as it depends on p and M. On the other hand, it does not depend on f. Given M and $p \in M$, there is a K such that *every* CR function on M is approximated uniformly on K by holomorphic polynomials.

Example 1.3.6 The K depends only on M, but can not always be all of M: For example, $M = S^1$ and $f = \bar{z}$.

The proof is based on the standard proof of Weierstraß theorem: If $f\colon [0,1] \to \mathbb{R}$ is continuous, then it is approximated on $[0,1]$ by the entire functions

$$f_\ell(z) = \int_0^1 c_\ell \, \mathrm{e}^{-\ell(z-t)^2} f(t)\, \mathrm{d}t \,, \tag{1.3.19}$$

for properly chosen c_ℓ. The proof involves taking partial sums of the power series. Baouendi–Trèves uses the same idea on a totally real subset of M and a slightly modified version of the above argument.

1.3.4 Extension of CR Functions

The following is called the Lewy extension theorem, but goes back to Helmut Knesser in 1936.

Theorem 1.3.4 *Suppose $M \subset \mathbb{C}^n$ is a smooth real hypersurface and $p \in M$. There exists a neighborhood U of p with the following property. Suppose $r\colon U \to \mathbb{R}$ is a smooth defining function for $M \cap U$, denote by $U_- \subset U$ the set where r is negative and $U_+ \subset U$ the set where r is positive. Let $f\colon M \to \mathbb{R}$ be a smooth CR function. Then:*

(i) If the Levi form with respect to r has a positive eigenvalue at p, then f extends to a holomorphic function on U_- continuous up to M.
(ii) If the Levi form with respect to r has a negative eigenvalue at p, then f extends to a holomorphic function on U_+ continuous up to M.
(iii) If the Levi form with respect to r has eigenvalues of both signs at p, then f extends to a holomorphic function on U.

So, if the Levi-form has eigenvalues of both signs, then every CR function is a restriction of a holomorphic function.

Let us provide a quick sketch of the proof of (i). First, suppose $p = 0$ and write M in the neighborhood of the origin as

$$\operatorname{Im} w = |z_1|^2 + \sum_{k=2}^{n-1} \epsilon_k |z_k|^2 + E(z_1, z', \bar{z}_1, \bar{z}', \operatorname{Re} w), \tag{1.3.20}$$

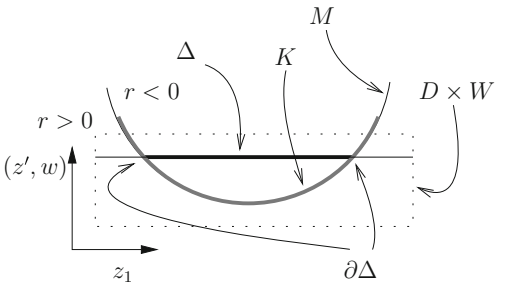
where $z' = (z_2, \ldots, z_{n-1})$, $\epsilon_k = -1, 0, 1$, and E is $O(3)$. Next, apply Bauoendi–Trèves theorem to find a compact neighborhood K of the origin of the theorem. The map

$$z_1 \mapsto |z_1|^2 + E(z_1, 0, \bar{z}_1, 0, 0)\,, \tag{1.3.21}$$

has a strict minimum at the origin, and so does

$$z_1 \mapsto |z_1|^2 + \sum_{k=2}^{n} \epsilon_k |z_k|^2 + E(z_1, z', \bar{z}_1, \bar{z}', \operatorname{Re} w) - \operatorname{Im} w \quad \text{for small } z', w. \tag{1.3.22}$$

We find an analytic disc Δ "attached" to $K \subset M$ (i.e., $\partial\Delta \subset K$):



(1.3.23)

One can fill a one-sided neighborhood by such discs.

The next step is to apply Baouendi–Trèves to find p_ℓ that approximate f uniformly on K. The sequence $\{p_\ell\}$ is (uniformly) Cauchy on $\partial\Delta$ for each disc. By the maximum principle, $\{p_\ell\}$ is (uniformly) Cauchy on Δ. This implies that $\{p_\ell\}$ is (uniformly) Cauchy on $U_- \cup K$, and it follows that $\{p_\ell\}$ converges to a holomorphic function on U_- continuous up to the boundary. To see (iii), extend to one side, then use the Tomato can principle or Kontinuitätssatz to extend to the other side.

Example 1.3.7 Every CR function on $\operatorname{Im} w = |z_1|^2 - |z_2|^2$ extends to an entire holomorphic function on $\mathbb{C}^3$ and hence must be real-analytic.

Example 1.3.8 Every CR function on $\operatorname{Im} w = |z_1|^2 + |z_2|^2$ extends to the set $\operatorname{Im} w \geqslant |z_1|^2 + |z_2|^2$, but not necessarily below.

Example 1.3.9 There exist CR functions on $\operatorname{Im} w = 0$ that extend to neither side.

Remark 1.3.2 These ideas led Lewy to find an example of an unsolvable PDE.

Another application is a special case of the following theorem.

Theorem 1.3.5 (Hartogs–Bochner) *Suppose $U \subset \mathbb{C}^n$, $n \geqslant 2$, is a bounded open set with a smooth boundary, and $f : \partial U \to \mathbb{C}$ is a CR function. Then there exists a continuous $F : \overline{U} \to \mathbb{C}$ holomorphic in U such that $F|_{\partial U} = f$.*

The special case is if we have at least one positive Levi eigenvalue at each point and if we can extend through compacts. This is something we will study in the next lecture.

Remark 1.3.3 Neither Hartogs nor Bochner proved this; it was proved by Martinelli.

Example 1.3.10 Every CR function on $S^{2n-1} \subset \mathbb{C}^n$, $n \geqslant 2$, is the boundary value of a continuous $F\colon \overline{\mathbb{B}_n} \to \mathbb{C}$ that is holomorphic in $\mathbb{B}_n$.

Example 1.3.11 The function $\bar{z}$ on $S^1 \subset \mathbb{C}$ is not the boundary value of a holomorphic function on the disc; it would have a pole.

Example 1.3.12 Similarly, this is not true in general if U is unbounded. If $U = \mathbb{D} \times \mathbb{C} \subset \mathbb{C}^2$, then $\bar{z}_1$ is a CR function but does not extend inside for the same reason.

1.4 The $\bar{\partial}$-Problem

Sean Curry

1.4.1 Levi Form

Let us return to the notions of pseudoconvexity that we started to explore in Sec. 1.2.

Recall that in the previous lectures we considered domains of $\mathbb{C}^n$ that can be written as $U = \{r < 0\}$, where $r = 0$ on the boundary $M = \partial U$. We considered the complexification of the tangent bundle, which takes the form

$$\mathbb{C}TM = T^{(1,0)}M \oplus T^{(0,1)}M \oplus \mathbb{C}T\,, \tag{1.4.1}$$

where $T_p^{(1,0)}M \oplus T_p^{(0,1)}M$ is the natural decomposition of $\mathbb{C}H_p$ into i and $-i$ egenspaces of J at each $p \in M$ and T is a choice of real tangent vector field that spans a complementary subspace[5] to H at each point. This decomposition of the complexified tangent space leads to a range of possible conditions on the Hessian of a defining function. As previously, let us write the Hessian along the tangent bundle as

$$\mathrm{Hess}(r)\big|_{TM} = \begin{pmatrix} r_{\alpha\beta} & r_{\alpha\bar{\beta}} & r_{\alpha 0} \\ r_{\bar{\alpha}\beta} & r_{\bar{\alpha}\bar{\beta}} & r_{\bar{\alpha}0} \\ r_{0\beta} & r_{0\bar{\beta}} & r_{00} \end{pmatrix}. \tag{1.4.2}$$

[5] This is sometimes referred to as "the bad direction." Note, however, that the article "the" is not totally justified in that, except for in certain cases of domains with special symmetry, there is no biholomorphically invariant way to choose this complementary subspace.

Then different the levels of convexity mentioned previously can be stated, informally, as follows:

$$\text{Convexity:} \qquad \operatorname{Hess}(r)|_{TM} \geqslant 0\,, \tag{1.4.3}$$

$$\mathbb{C}\text{-linear convexity:} \qquad \begin{pmatrix} r_{\alpha\beta} & r_{\alpha\bar{\beta}} \\ r_{\bar{\alpha}\beta} & r_{\bar{\alpha}\bar{\beta}} \end{pmatrix} \geqslant 0\,, \tag{1.4.4}$$

$$\text{Pseudoconvexity:} \qquad r_{\alpha\bar{\beta}} \geqslant 0\,. \tag{1.4.5}$$

More precisely, by (1.4.3) we mean that the Hessian of r is positive semidefinite as a quadratic form at each point $p \in M$ when restricted to real tangent vectors to M, by (1.4.4) we mean that the Hessian of r is positive semidefinite as a quadratic form at each point $p \in M$ when restricted to $H_p \subseteq T_pM$ (strictly speaking, since $T^{(1,0)}M \oplus T^{(0,1)}M = \mathbb{C}H$, the notation in (1.4.4) translates to "$\operatorname{Hess}(r)|_{\mathbb{C}H} \geqslant 0$," but what we really mean is that $\operatorname{Hess}(r)|_H \geqslant 0$, i.e., the corresponding *real* quadratic form is positive semidefinite), and by (1.4.5) we mean that the Hessian of r defines a positive semidefinite Hermitian form on $T^{(1,0)}_p M$ for each $p \in M$, that is, at each point p we have $\operatorname{Hess}(r)(X_p, \overline{X_p}) \geqslant 0$ for all $X_p \in T^{(1,0)}_p M$.

Let us focus on pseudoconvexity, which is the weakest requirement (and the only one that is biholomorphically invariant). In practice, it amounts to checking that

$$\sum_{j,k=1}^{n} a_j \bar{a}_k \frac{\partial^2 r}{\partial z_j \partial \bar{z}_k} \geqslant 0 \qquad \text{for all} \quad X_p = \sum_{j=1}^{n} a_j \frac{\partial r}{\partial z_j}\bigg|_p \in T^{(1,0)}_p M\,. \tag{1.4.6}$$

The Hermitian quadratic form on the left-hand side is called the *Levi form* of r (or of $M = \partial U$):

$$\operatorname{Levi}(r)|_p(X_p, Y_p) = \operatorname{Hess}(r)|_p(X_p, \overline{Y_p}) \qquad \text{for} \quad X_p, Y_p \in T^{(1,0)}_p M. \tag{1.4.7}$$

Pseudoconvexity just says that $\operatorname{Levi}(r) \geqslant 0$ as a Hermitian form. The signature of the Levi form is invariant and well-defined (it is independent of the choice of defining function and is invariant under local biholomorphic changes of coordinates). This is a consequence of the fact that the Levi form is actually invariant as a line bundle-valued Hermitian form on $T^{(1,0)}M$.

Example 1.4.1 Consider $\mathbb{C}^3$ with complex Cartesian coordinates (z_1, z_2, w). The domain defined by $\operatorname{Im} w > |z_1|^2 + |z_2|^2$ has boundary M defined by $\operatorname{Im} w = |z_1|^2 + |z_2|^2$. A natural choice of defining function for our domain is $r = |z_1|^2 + |z_2|^2 - \operatorname{Im} w$. A frame for the holomorphic tangent bundle $T^{(1,0)}M$ is given by

$$Z_1 = \frac{\partial}{\partial z_1} + 2i\bar{z}_1 \frac{\partial}{\partial w}, \qquad Z_2 = \frac{\partial}{\partial z_2} + 2i\bar{z}_2 \frac{\partial}{\partial w}. \tag{1.4.8}$$

To see this it suffices to note that Z_1 and Z_2 are pointwise linearly independent, lie in the span of the holomorphic vector fields $\frac{\partial}{\partial z_1}$, $\frac{\partial}{\partial z_2}$ and $\frac{\partial}{\partial w}$ on $\mathbb{C}^3$ and satisfy $Z_\alpha r = 0$ for $\alpha = 1, 2$ (these vector fields happen to be defined and satisfy these conditions on $\mathbb{C}^3$ but we really only need these things to be satisfied on M). A frame for $T^{(1,0)}M$ is given by $Z_{\bar{1}} = \overline{Z_1}$, $Z_{\bar{2}} = \overline{Z_2}$. We wish to compute the Levi form components $\mathrm{Hess}(r)(Z_\alpha, Z_{\bar{\beta}})$. Note that for general (real or complex) tangent vector fields X and Y on $\mathbb{C}^3$, computing $\mathrm{Hess}(r)(X, \overline{Y})$ is not the same as computing $X(Yr)$ or $Y(Xr)$ since $\mathrm{Hess}(r)(X, \overline{Y})$ is tensorial in X and Y whereas the expression $X(Yr)$ involves the X derivatives of the coefficients of Y and $Y(Xr)$ involves the X derivatives of the coefficients of Y; the correct formula, for arbitrary vector fields X and Y, is

$$\mathrm{Hess}(r)(X, Y) = Y(Xr) - (\nabla_Y X)r, \tag{1.4.9}$$

where $\nabla_Y X$ is the vector field obtained from X by differentiating its components in the direction of Y and all vector fields on the right hand side of the above display are acting on the functions to their right as directional derivative operators. Using the above displayed formula it is straightforward to compute that the Levi form of r has components

$$\mathrm{Hess}(r)(Z_\alpha, Z_{\bar{\beta}}) = 2\delta_{\alpha\bar{\beta}}\,, \tag{1.4.10}$$

at any point $p \in M$ (in more general examples the components of the Levi form expressed in a general frame will be functions rather than constants; but, if the Levi form is positive definite at some point, then by Gram-Schmidt it is possible to find a frame in a neighborhood of that point such that the Levi form has components $\delta_{\alpha\bar{\beta}}$, or $2\delta_{\alpha\bar{\beta}}$ if preferred, at all points in that neighborhood).

Example 1.4.2 Again, consider $\mathbb{C}^3$ with coordinates (z_1, z_2, w). The domain defined by $\mathrm{Im}\ w > |z_1|^2 - |z_2|^2$ has boundary M defined by $\mathrm{Im}\ w = |z_1|^2 - |z_2|^2$. A frame for the holomorphic tangent bundle $T^{(1,0)}M$ is given by

$$Z_1 = \frac{\partial}{\partial z_1} + 2i\bar{z}_1\frac{\partial}{\partial w}, \qquad Z_2 = \frac{\partial}{\partial z_2} - 2i\bar{z}_2\frac{\partial}{\partial w}, \tag{1.4.11}$$

and a frame for $T^{(1,0)}M$ is given by $Z_{\bar{1}} = \overline{Z_1}$, $Z_{\bar{2}} = \overline{Z_2}$. Computing as in the previous example, the Levi form of the defining function $r = |z_1|^2 - |z_2|^2 - \mathrm{Im}\ w$ is represented in the frame Z_1, Z_2 by the matrix $\mathrm{diag}\{2, -2\}$.

Example 1.4.3 Consider $\mathbb{C}^n$ with the coordinates $(z, w) = (z_1, \ldots, z_{n-1}, w)$. The domain defined by $\mathrm{Im}\ w > 0$ has boundary M defined by $\mathrm{Im}\ w = 0$. A frame for the holomorphic tangent bundle $T^{(1,0)}M$ is given by the vector fields $Z_\alpha = \frac{\partial}{\partial z_\alpha}$ for $\alpha = 1, \ldots, n-1$. The defining function $r = -\mathrm{Im}\ w$ is linear, meaning that $\mathrm{Hess}(r) \equiv 0$, and hence the Levi form of this domain is identically zero; we say that

its boundary is *Levi flat* . Note that the boundary$M = \mathbb{C}^{n-1} \times \mathbb{R}$, being Levi flat, is pseudoconvex from both sides (meaning that both $\operatorname{Im} w > 0$ and $\operatorname{Im} w < 0$ are pseudoconvex domains).

The hypersurfaces $\operatorname{Im} w = \epsilon_1|z_1|^2 + \cdots + \epsilon_{n-1}|z_{n-1}|^2$, where $\epsilon_k \in \{1, -1, 0\}$, are local models for general hypersurfaces (when viewed up to local biholomorphism) in the following sense (Lem. 2.3.9 in [1]):

Proposition 1.4.1 *Let $M \subset \mathbb{C}^n$ be a smooth real hypersurface and $p \in M$. Then there exists a local biholomorphic change of coordinates taking p to 0 and M to the hypersurface given by*

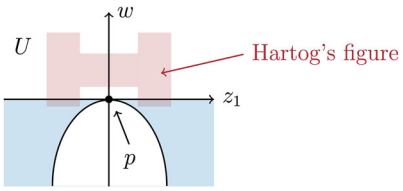
$$Im\, w = \sum_{k=1}^{\alpha} |z_k|^2 - \sum_{k=\alpha+1}^{\alpha+\beta} |z_k|^2 + E(z, \bar{z}, Re\, w), \tag{1.4.12}$$

where E is $O(3)$ at the origin. Here α is the number of positive eigenvalues of the Levi form at p, β is the number of negative eigenvalues and $\alpha + \beta \leqslant n - 1$.

By constructing a suitable Hartogs figure (or "tomato can") it is easy to show that holomorphic functions on $\operatorname{Im} w > |z_1|^2 - |z_2|^2$ can be extended to a neighborhood of the origin (in fact, to all of $\mathbb{C}^3$, but we focus here on local extendability). Combining this observation (and its natural generalization to higher dimensions) with Proposition 1.4.1 one readily obtains the following theorem (Thm. 2.3.11 in [1]):

Theorem 1.4.1 (Tomato Can Principle) *Suppose $U \subset \mathbb{C}^n$ is an open set with smooth boundary and the Levi form has a negative eigenvalue at $p \in \partial U$. Then every holomorphic function on U extends to a neighborhood of p. In particular, U is not a domain of holomorphy.*

The following figure illustrates the idea behind the proof, where here we have applied Proposition 1.4.1 and permuted the z_k variables so that z_1 is a direction in which the Levi form is negative at p.



(1.4.13)

From the above theorem it follows that a domain of holomorphy must be pseudoconvex (since this is equivalent to the eigenvalues of the Levi form being nonnegative at every point).

1.4.2 Equivalences

We conclude our discussion of domains of holomorphy in $\mathbb{C}^n$ with the following theorem, giving a characterization of such domains in terms of pseudoconvexity (and some final examples). At this point we allow for domains with non-smooth boundary, with the caveat that we must then work with a different notion of pseudoconvexity (due to Hartogs) that makes sense when the boundary is not smooth and is equivalent to the (Levi) definition we have been using in the case of smooth boundary. A longer list of conditions equivalent to being a domain of holomorphy can be found in [1] (see also [11]).

Theorem 1.4.2 *Let $U \subseteq \mathbb{C}^n$ be a domain. Then the following are equivalent:*

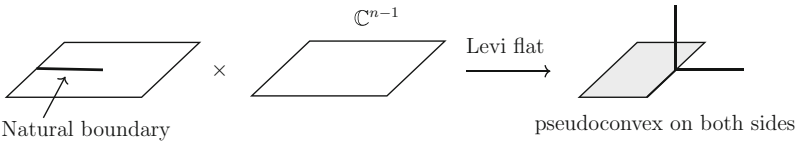
(i) U is a domain of holomorphy;
(ii) U is Levi pseudoconvex (provided ∂U is smooth);
(iii) U is Hartogs pseudoconvex, i.e., there is a continuous exhaustion function that is plurisubharmonic.

Here, an exhaustion function is one that goes to infinity at the boundary. Plurisubharmonic means subharmonic on each complex line (i.e., on each one dimensional complex affine subspace of $\mathbb{C}^n$). It is essentially a positivity condition on the complex Hessian; morally it means "$\partial^2 f/\partial z_j \partial \bar{z}_k \geqslant 0$" (it means precisely this if f is twice continuously differentiable, but the condition still makes sense if f is merely continuous).

We have only proved that (i) implies (ii) in the case where ∂U is smooth. That the notions of Levi pseudoconvexity and Hartogs pseudoconvexity are equivalent when ∂U is smooth is proved, e.g., in [1, Thm. 2.5.8]. (This equivalence allows us to say that a domain is *pseudoconvex* without needing to specify which of the two definitions we are using.) That (i) implies (iii) is typically seen by showing that the domains of holomorphy are "holomorphically convex" (see [1, Thm. 2.6.3] or [11, Thm. 2.5.4 & 2.5.5]) and hence that minus the logarithm of the distance to the boundary is purisubharmonic (see [11, pp. 44–45]). Proving that a pseudoconvex domain is a domain of holomorphy is known as the Levi problem (as it was posed, at least for the smooth boundary case, by Levi in 1911) and was solved in the early 1950s due, principally, to the work of K. Oka. This problem played an important role in the development of analysis in several complex variables. Today, the solution to the Levi problem is most commonly presented using the techniques for the $\overline{\partial}$-problem developed by Kohn, Hörmander and Andreotti-Vesentini (and others) in the mid 1960s; see [11].

Although the prototypical example of a pseudoconvex domain in $\mathbb{C}^n$ would be either the unit ball or the unit polydisc (both are convex, and hence pseudoconvex), in physics-related problems it seems that one might more commonly encounter domains with Levi flat boundaries, such as in the following example.

Example 1.4.4 An example that one might see in the physics literature is the domain $U = \mathbb{C}^n \setminus (\mathbb{R}_{\leqslant 0} \times \mathbb{C}^{n-1}) \subseteq \mathbb{C}^n$ (imagine having a holomorphic function $f(z_1, z_2 \ldots, z_n)$ that is entire in each of the variables $z_2, \ldots, z_n$ but has a natural boundary along the negative real axis in the z_1 variable). Note that the boundary is then $M = \mathbb{R}_{\leqslant 0} \times \mathbb{C}^{n-1}$, which (up to a permutation of the coordinates) is a part of the boundary considered in Example 1.4.3 and so its interior $\mathbb{R}_{\leqslant 0} \times \mathbb{C}^{n-1}$ is Levi flat (pseudoconvex from both sides). That the domain U is pseudoconvex can be seen as coming from the fact that all domains in $\mathbb{C}$ are pseudoconvex, since U is the product of the cut plane $\mathbb{C} \setminus \mathbb{R}_{\leqslant 0}$ with $\mathbb{C}^{n-1}$ and the product of pseudoconvex domains is pseudoconvex (i.e., a domain of holomorphy).



(1.4.14)

We have talked a lot about CR manifolds embedded as hypersurfaces in complex Euclidean space, but (just as we can abstract the notion of a surface in Euclidean space to get the notion of an abstract Riemannian manifold) one can also talk about abstract CR manifolds. These are $2n + 1$ dimensional real smooth manifolds M equipped with a complex rank n subbundle $T^{(1,0)}M$ of the complexified tangent bundle $\mathbb{C}TM$ with the property that $T^{(1,0)}_p M \cap \overline{T^{(1,0)}_p M} = \{0\} \subseteq \mathbb{C}T_p M$ for all $p \in M$ and such that for any two vector fields $X, Y \in \Gamma(T^{(1,0)}M)$ the Lie bracket $[X, Y]$ also lies in $T^{(1,0)}M$. The CR structure $T^{(1,0)}M$ of any smooth real hypersurface in $\mathbb{C}^{n+1}$ satisfies these conditions.[6] Here is an example of an abstractly defined CR manifold (which will turn out to be equivalent to a familiar hypersurface in $\mathbb{C}^{n+1}$):

Example 1.4.5 Consider the abstract Heisenberg CR structure on $\mathbb{H}_n = \mathbb{C}^n \times \mathbb{R}$ (not viewed as a subset of $\mathbb{C}^{n+1}$) with coordinates $(z_1, \ldots, z_n, t)$. We have

$$T^{1,0}\mathbb{H}_n = \operatorname{span}\{Z_1, Z_2, \ldots, Z_n\}, \tag{1.4.15}$$

where

$$Z_j = \frac{\partial}{\partial z_j} + i\bar{z}_j \frac{\partial}{\partial t}. \tag{1.4.16}$$

[6] Some people prefer to talk about hypersurfaces in $\mathbb{C}^{n+1}$ rather than in $\mathbb{C}^n$ so that the dimension of the hypersurface is $2n+1$ rather than $2n-1$ and the CR dimension is n rather than $n-1$. Sean is doing this locally (by force of habit) for this example. There is no other significance in this change of convention.

Let us write $Z_{\bar{j}} = \overline{Z_j}$ and $T = \frac{\partial}{\partial t}$. Then, one can show that

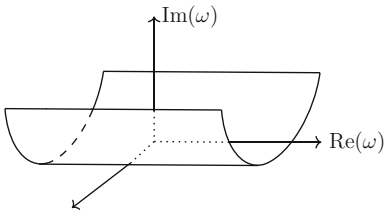
$$[T, Z_j] = 0\,, \qquad [Z_j, Z_k] = 0\,, \qquad [Z_j, \bar{Z}_k] = -2i\delta_{j\bar{k}}T\,. \tag{1.4.17}$$

Here, $2\delta_{j\bar{k}}$ are the abstract Levi form components (the Levi form can be defined abstractly as $i[X, \overline{Y}] \mod T^{(1,0)} \oplus T^{(1,0)}$ for sections X and Y of $T^{(1,0)}$; this expression is easily checked to be tensorial in X and Y and gives a well-defined line bundle valued Hermitian form on $T^{(1,0)}$, where the line bundle is the complex tangent bundle mod $T^{(1,0)} \oplus T^{(1,0)}$; one needs to trivialize this line bundle to get an ordinary Hermitian form on $T^{(1,0)}$ and this is done using the choice of T which is analogous to the choice of defining function r in the embedded case). The Heisenberg CR structure gets its name from these commutation relations and the (corresponding) fact that there is a natural Heisenberg group structure on $\mathbb{H}_n = \mathbb{C}^n \times \mathbb{R}$ with respect to which the frame $Z_1, Z_2, \ldots, Z_n$ is left invariant.

The map

$$(z_1, \ldots, z_n, t) \mapsto (z_1, \ldots, z_n, \underbrace{t + i\|z\|^2}_{w})\,, \tag{1.4.18}$$

embeds the abstract CR manifold $(\mathbb{H}_n, T^{(1,0)}\mathbb{H}_n)$ as the boundary M of $\Omega_0 = \{\operatorname{Im} w > \|z\|^2\}$; this map is a *CR embedding* in the sense that it identifies the abstract CR structure $(\mathbb{H}_n, T^{(1,0)}\mathbb{H}_n)$ with the embedded one $(M, T^{(1,0)}M)$, meaning that the map is a smooth embedding with image M and the differential of the embedding at each point $p \in \mathbb{H}_n$ induces an isomorphism between $T^{(1,0)}_p\mathbb{H}_n$ and $T^{(1,0)}_p M$. A smooth embedding is a CR embedding if and only if its component functions satisfy the *tangential Cauchy-Riemann equations*, meaning that the functions are in the kernel of some, equivalently any, local frame $Z_{\bar{1}}, \ldots, Z_{\bar{n}}$ for the antiholomorphic tangent bundle of the abstract CR manifold in a neighborhood of each point. In the case of our example it is easy to check that the components $z_1, \ldots, z_n, w = t + i\|z\|^2$ of the map (1.4.18) satisfy the tangential Cauchy-Riemann equations (i.e., lie in the kernel of the operators $Z_{\bar{1}}, \ldots, Z_{\bar{n}}$) and this is the quickest way to see that the map is a CR embedding.



(1.4.19)

Note that in this example, the abstract holomorphic tangent vector fields $Z_1, \ldots, Z_n$ defined in (1.4.16) are pushed forward by the diffeomorphism $\mathbb{H}_n \to M \subseteq \mathbb{C}^{n+1}$ to the vector fields

$$\frac{\partial}{\partial z_1} + 2i\bar{z}_1 \frac{\partial}{\partial w}\,, \quad \ldots \quad , \quad \frac{\partial}{\partial z_n} + 2i\bar{z}_n \frac{\partial}{\partial w}\,, \tag{1.4.20}$$

defined along M, where in this last display $z_1, \ldots z_n$ are the first n coordinates on $\mathbb{C}^{n+1}$ whereas in (1.4.16) the $z_1, \ldots z_n$ are coordinates for the $\mathbb{C}^n$ factor of the abstract manifold $\mathbb{H}_n = \mathbb{C}^n \times \mathbb{R}$ (so "$\frac{\partial}{\partial z_j}$" does not mean the same thing in (1.4.16) as it does in (1.4.20); it might have been wise to use $(\zeta_1, \ldots, \zeta_n, w)$ rather than $(z_1, \ldots, z_n, w)$ to denote the coordinates on $\mathbb{C}^{n+1}$ but we chose not to). To see this, let $u = \operatorname{Re} w$ and $v = \operatorname{Im} w$ and note that $\frac{\partial}{\partial w} = \frac{1}{2}\left(\frac{\partial}{\partial u} - i\frac{\partial}{\partial v}\right)$. Clearly the vector field $\frac{\partial}{\partial t}$ on $\mathbb{H}_n$ is pushed forward to the vector field $\frac{\partial}{\partial u}$ on M. On the other hand, since w also depends on z, the vector field $\frac{\partial}{\partial z_j}$ on $\mathbb{H}_n$ is pushed forward to the vector field $\frac{\partial}{\partial z_j} + i\bar{z}_j\frac{\partial}{\partial w} - i\bar{z}_j\frac{\partial}{\partial \bar{w}} = \frac{\partial}{\partial z_j} + \bar{z}_j\frac{\partial}{\partial v}$ (a quick way to check this is to use $(\zeta_1, \ldots, \zeta_n, w)$ instead of $(z_1, \ldots, z_n, w)$ for the coordinates on $\mathbb{C}^{n+1}$, so that $\zeta_j = z_j$ and $w = t + i\|z\|^2$, and then use the chain rule to relate the various partial derivative operators, where one thinks of the map $\mathbb{H}_n \to M \subseteq \mathbb{C}^{n+1}$ as a parametrization of M by (z, t)). Combining these observations establishes our claim (since $i\bar{z}_j\frac{\partial}{\partial u} + \bar{z}_j\frac{\partial}{\partial v} = 2i\bar{z}_j\frac{\partial}{\partial w}$).

Finally, note that $r = \|z\|^2 - \operatorname{Im} w$ is a natural choice of defining function for Ω_0 and (as in Example 1.4.1) one can easily verify that the Levi form of r in this frame has components $2\delta_{j\bar{k}}$. In particular, the Levi form is positive definite at every point.

1.4.3 Introduction to the $\bar{\partial}$-Equation

First of all, we need to understand what is the $\bar{\partial}$-equation ($\bar{\partial}$ is pronounced "dee bar"). Let us introduce the notation

$$\bar{\partial} f = \left(\frac{\partial f}{\partial \bar{z}_1}, \ldots, \frac{\partial f}{\partial \bar{z}_n}\right) = f_{\bar{z}_1} \mathrm{d}\bar{z}_1 + \ldots + f_{\bar{z}_n} \mathrm{d}\bar{z}_n\,. \tag{1.4.21}$$

It will be the most convenient to think of $\bar{\partial} f$ as a $(0, 1)$-form, as on the right-hand side. We assume some familiarity with forms, though it will not be essential. On $U \subseteq \mathbb{C}^n$, the $\bar{\partial}$ operator on functions described above extends to a family of maps between the spaces $\wedge^{0,k}$ of $(0, k)$-forms:

$$\wedge^{0,0} \xrightarrow{\bar{\partial}} \wedge^{0,1} \xrightarrow{\bar{\partial}} \wedge^{0,2} \xrightarrow{\bar{\partial}} \ldots \xrightarrow{\bar{\partial}} \wedge^{0,n}\,. \tag{1.4.22}$$

The space $\wedge^{0,n}$, for example, is spanned by forms proportional to $\mathrm{d}\bar{z}_1 \wedge \cdots \wedge \mathrm{d}\bar{z}_n$. We will focus our attention on the first operator $\wedge^{0,0} \xrightarrow{\bar{\partial}} \wedge^{0,1}$ defined as in (1.4.21), though this necessitates some consideration of the next operator $\wedge^{0,1} \xrightarrow{\bar{\partial}} \wedge^{0,2}$.

For a $(0, 1)$-form g, we write $g = g_1 \mathrm{d}\bar{z}_1 + \ldots + g_n \mathrm{d}\bar{z}_n$. Then

$$\bar{\partial} g = 0 \qquad \Leftrightarrow \qquad \partial_{\bar{z}_j} g_{\bar{k}} = \partial_{\bar{z}_k} g_{\bar{j}} \quad \text{for all } j, k. \tag{1.4.23}$$

Finally, it is straightforward to show that since the second partial derivatives commute we have

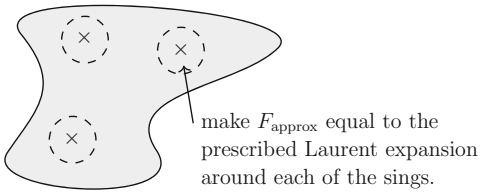
$$\bar{\partial}^2 = 0\,. \tag{1.4.24}$$

We are now ready to state the $\bar{\partial}$-problem (in the lowest degree case): Given a $(0, 1)$-form g satisfying the obvious necessary condition (1.4.23), can we solve $\bar{\partial} f = g$?

A sharper question is: can we find a solution operator? Alternatively, can we "solve with estimates"? The last class of problems has for a long time been one of the most active areas of research in several complex variables, tracing its roots back to the work of Kohn, Hörmander and others in the 1960s. Being able to "solve with estimates" is very useful. Suppose that we are trying to construct a holomorphic function F (or some other "holomorphic" object) with certain desired properties, and we know that we can construct such a function that is smooth and close to being holomorphic (in an appropriate sense). Then we can take our approximately holomorphic function F_{approx} and set $g = \bar{\partial} F_{\text{approx}}$, which measures the failure of our approximate solution to be holomorphic and should be "small" (and will satisfy $\bar{\partial} g = 0$ since $\bar{\partial}^2 = 0$). Then, if we are able to solve the $\bar{\partial}$-problem with estimates, this gives a solution f to $\bar{\partial} f = g$ with f "small" (in some norm it should be bounded by some other norm of g, which is small). In particular, the smallness of f should force f to be different from, and relatively small compared to F_{approx} (both solve the $\bar{\partial}$-equation with g as the right-hand side, but remember that the space of solutions to this equation is very large since holomorphic functions lie in the kernel of $\bar{\partial}$); this is done by choosing appropriate function spaces/norms. Our desired holomorphic function F is then $F_{\text{approx}} - f$, and the "smallness" of f ensures that F retains the desired characteristics that F_{approx} had by construction.

As a simple example of this (where the only "estimates" we need are the local boundedness of the solution f) suppose that we are supposed to find a meromorphic function F on $\mathbb{C}$ with prescribed poles at a finite number of points $p_1, \ldots, p_m$ and with F having prescribed principal part of its Laurent expansion (the part involving negative powers) at each of these poles. Picking discs $\Delta_{r_1}(p_1), \ldots, \Delta_{r_m}(p_m)$ around these points whose closures are disjoint, we may define F_{approx} in each punctured disc $\Delta_{r_j}(p_j) \setminus \{p_j\}$ to be equal to the polynomial in $\frac{1}{z-p_j}$ that is prescribed as the principal part of the Laurent expansion of F. We may then smoothly extend F_{approx} to all of $\mathbb{C}$ with compact support (meaning that F_{approx}

is zero outside of a large enough ball).



(1.4.25)

Note that $g = \bar{\partial} F_{\text{approx}}$ will be zero inside each of the punctured discs, and hence can be extended by zero across the points $p_1, \ldots, p_m$ so as to be defined on all of $\mathbb{C}$. Thus, we have a compactly supported smooth $(0,1)$-form g (that trivially satisfies $\bar{\partial} g = 0$) and if we know that there is a locally bounded solution f to $\bar{\partial} f = g$, then $F = F_{\text{approx}} - f$ will be holomorphic on $\mathbb{C} \setminus \{p_1, \ldots, p_m\}$ and (due to the boundedness of f near each point p_j) will have the prescribed polynomial in $\frac{1}{z-p_j}$ as the principle part of its Laurent expansion at each of the p_j. We will see in the discussion that follows how to find such a function f.

1.4.4 The Generalized Cauchy Integral Formula

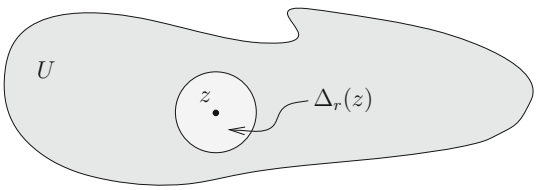
Before we get into the $\bar{\partial}$-problem, let us state a more general version of Cauchy's formula using Stokes' theorem (really, Green's theorem). This version is called the Cauchy–Pompeiu integral formula. We only need the theorem for smooth functions, but, as it is often applied in less regular contexts and is just an application of Stokes' theorem, let us state it more generally. In applications, the boundary is often only piecewise smooth, and again that is all we need for Stokes.

Theorem 1.4.3 *In one complex variable, we have the Cauchy–Pompeiu formula:*

$$f(z) = \frac{1}{2\pi i} \int_{\partial U} \frac{f(\zeta)}{\zeta - z} d\zeta + \frac{1}{2\pi i} \int_U \frac{\frac{\partial f}{\partial \bar{\zeta}}(\zeta)}{\zeta - z} \underbrace{d\zeta \wedge d\bar{\zeta}}_{-2i\,dx \wedge dy} . \tag{1.4.26}$$

for $z \in U$ with piecewise C^1 boundary ∂U and $f : \bar{U} \to \mathbb{C}$ continuous with bounded partial derivatives in U.

The proof involves excising a small disk $\Delta_r(z)$ from U and applying Stokes' theorem on $U \setminus \Delta_r(z)$.



(1.4.27)

The basic idea for several complex variables is to recycle the above 1D problem but also use $\bar\partial g = 0$, which means that the derivatives are not independent of each other and hence allows us to tie different 1D problems together.

1.4.5 Compactly Supported $\bar\partial$-Problem

Let us now consider $U = \mathbb{C}^n$ with $n \geqslant 2$. Most of what we say below will also work for $n = 1$.

Theorem 1.4.4 *Suppose g is a $(0,1)$-form on $\mathbb{C}^n$ with $n \geqslant 2$, which is smooth, compactly supported, and satisfies the integrability condition $\bar\partial g = 0$ ($\partial_{\bar z_j} g_k = \partial_{\bar z_k} g_j$ for all j, k) on $\mathbb{C}^n$. Then there exists a unique compactly supported smooth function $\psi : \mathbb{C}^n \to \mathbb{C}$ such that*

$$\bar\partial \psi = g\,, \tag{1.4.28}$$

i.e., $\partial\psi/\partial\bar z_j = g_j$ for all $j = 1, 2, \ldots, n$.

Proof We are motivated by the Cauchy–Pompeiu formula with g_1 replaced by $\partial\psi/\partial\bar z_1$ and U a "large" ball. We will define ψ by

$$\psi(z) = \frac{1}{2\pi i}\int_{\mathbb{C}} \frac{g_1(\zeta, z_2, \ldots, z_n)}{\zeta - z_1} \mathrm{d}\zeta \wedge \mathrm{d}\bar\zeta\,. \tag{1.4.29}$$

The aim is to differentiate the above formula and check that it satisfies (1.4.28). The non-trivial task is going to be to establish that the result is compactly supported. We first change the variables $\zeta \mapsto \zeta + z_1$, which give

$$\psi(z) = \frac{1}{2\pi i}\int_{\mathbb{C}} \frac{g_1(\zeta + z_1, z_2, \ldots, z_n)}{\zeta} \mathrm{d}\zeta \wedge \mathrm{d}\bar\zeta\,. \tag{1.4.30}$$

This eliminates z_1 from the denominator, so we can take the derivative $\partial/\partial\bar z_k$ under the integral. Hence

$$\frac{\partial\psi}{\partial\bar z_k}(z) = \frac{1}{2\pi i}\int_{\mathbb{C}} \frac{\frac{\partial g_1}{\partial\bar z_k}(\zeta + z_1, z_2, \ldots, z_n)}{\zeta} \mathrm{d}\zeta \wedge \mathrm{d}\bar\zeta\,. \tag{1.4.31}$$

On the other hand, the Cauchy–Pompeiu formula applied to g_k (on $|z| \leqslant R$ for large R) gives

$$g_k(z) = \frac{1}{2\pi i}\int_{\mathbb{C}} \frac{\frac{\partial g_k}{\partial\bar z_1}(\zeta, z_2, \ldots, z_n)}{\zeta - z_1} \mathrm{d}\zeta \wedge \mathrm{d}\bar\zeta\,. \tag{1.4.32}$$

We can now use the integrability condition $\partial g_1/\partial \bar{z}_k = \partial g_k/\partial \bar{z}_1$ to write

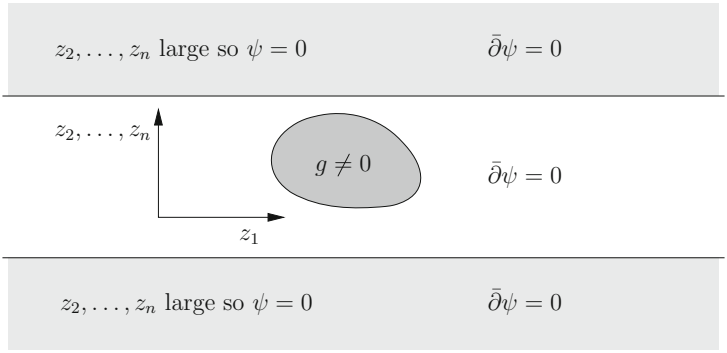
$$\frac{\partial \psi}{\partial \bar{z}_k}(z) = \frac{1}{2\pi i} \int_{\mathbb{C}} \frac{\frac{\partial g_k}{\partial \bar{z}_1}(\zeta + z_1, z_2, \ldots, z_n)}{\zeta} \mathrm{d}\zeta \wedge \mathrm{d}\bar{\zeta} \,. \tag{1.4.33}$$

Finally, making the change of variables $\zeta \mapsto \zeta - z_1$, we find

$$\frac{\partial \psi}{\partial \bar{z}_k}(z) = \frac{1}{2\pi i} \int_{\mathbb{C}} \frac{\frac{\partial g_k}{\partial \bar{z}_1}(\zeta, z_2, \ldots, z_n)}{\zeta - z_1} \mathrm{d}\zeta \wedge \mathrm{d}\bar{\zeta} = g_k(z) \,, \tag{1.4.34}$$

by the previous formula (1.4.31).

So far, the proof works for $n \geqslant 1$. It remains to establish that ψ has a compact support, which requires $n \geqslant 2$. Consider the picture:



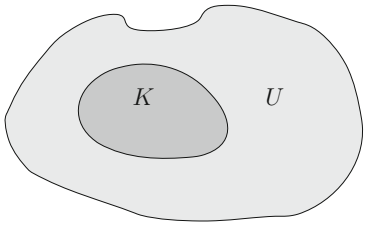
(1.4.35)

From the formula for ψ, $\psi \equiv 0$ in the regions where at least one of $z_2, \ldots, z_n$ is large. However, by the identity theorem, this forces $\psi \equiv 0$ in the unbounded component of $\mathbb{C}^n \setminus \mathrm{supp}(g)$. Thus, ψ has compact support. □

1.4.6 The General Hartogs Phenomenon

We can now prove the general Hartogs phenomenon as an application of the solution of the compactly supported inhomogeneous $\bar{\partial}$-problem. We proved special versions of this phenomenon using Hartogs figures before.

Theorem 1.4.5 (Hartogs Phenomenon) *Let $U \subseteq \mathbb{C}^n$ be a domain with $n \geqslant 2$, and let $K \subseteq U$ be a relatively compact subset of U such that $U \setminus K$ is connected. Then every holomorphic function $f \in \mathcal{O}(U \setminus K)$ extends holomorphically to all of U.*



(1.4.36)

The idea behind the proof is as follows. First, extend $f \in \mathcal{O}(U \setminus K)$ smoothly to get the (not necessarily holomorphic) function $\tilde{f}$ on U, with $\bar{\partial}\tilde{f}$ supported in K (one may need to enlarge K slightly here if it is not a nice set with smooth boundary as in the picture, but this is not a problem). Then, solve $\bar{\partial}\psi = \bar{\partial}\tilde{f}$ with ψ compactly supported in K. Finally, set $f = \tilde{f} - \psi$ on U.

We can finally add yet another equivalence statement to the list in Thm. 1.4.2:
(iv) The $\bar{\partial}$-problem ($\bar{\partial} f = g$) is solvable on $(0, q)$-forms for $1 \leqslant q \leqslant n - 1$.

This discussion illustrates the general idea of how one uses the solvability of the $\bar{\partial}$-problem. There are a whole myriad of other setups where we can solve the $\bar{\partial}$-problem. For example, it is solvable on polydiscs, balls, and more generally on domains of holomorphy, and so in these cases one is naturally lead to ask about "solvability with estimates" (the classic references are [11, 14], more recent references include [15–17]). Solvability of the $\bar{\partial}$-problem also plays a key role in complex geometry; see, e.g., [18]. However, for the purposes of these lectures, we will stop here.

References

1. J. Lebl, Tasty bits of several complex variables (Independently Published, 2023). https://www.jirka.org/scv/
2. C. Fefferman, The Bergman kernel and biholomorphic mappings of pseudoconvex domains. Invent. Math. **26**, 1 (1974)
3. C.L. Fefferman, Monge-Ampère equations, the Bergman kernel, and geometry of pseudoconvex domains. Ann. Math. **103**, 395 (1976)
4. C. Fefferman, Parabolic invariant theory in complex analysis. Adv. Math. **31**, 131 (1979)
5. C. Fefferman, C.R. Graham, Conformal invariants, in *Élie Cartan et les Mathématiques d'aujourd'hui - Lyon, 25–29 Juin 1984* (Société mathématique de France, Paris, 1985), pp. 95–116
6. C. Fefferman, C.R. Graham, *The Ambient Metric*. Annals of Mathematics Studies, vol. 178 (Princeton University Press, Princeton, 2012)
7. I. Robinson, A. Trautman, Some spherical gravitational waves in general relativity. Proc. Roy. Soc. Lond. Ser. A **265**, 463 (1961–1962)
8. R.P. Kerr, Gravitational field of a spinning mass as an example of algebraically special metrics. Phys. Rev. Lett. **11**, 237 (1963)
9. R. Penrose, W. Rindler, *Spinors and Space-Time*. Cambridge Monographs on Mathematical Physics, vol. 2, 2nd edn. (Cambridge University Press, Cambridge, 1988)
10. C.D. Hill, J. Lewandowski, P. Nurowski, Einstein's equations and the embedding of 3-dimensional CR manifolds. Indiana Univ. Math. J. **57**, 3131 (2008)

11. L. Hormander, *An Introduction to Complex Analysis in Several Variables* (Elsevier, Amsterdam, 1973)
12. J.J. Kohn, L. Nirenberg, A pseudo-convex domain not admitting a holomorphic support function. Math. Ann. **201**, 265 (1973)
13. M. Kolář, Higher order invariants of Levi degenerate hypersurfaces. Pure Appl. Math. Q. **6**, 1035 (2010)
14. G.B. Folland, J.J. Kohn, *The Neumann Problem for the Cauchy-Riemann Complex*. Annals of Mathematics Studies, vol. 75 (Princeton University Press, Princeton/University of Tokyo Press, Tokyo, 1972)
15. S.-C. Chen, M.-C. Shaw, *Partial Differential Equations in Several Complex Variables*. AMS/IP Studies in Advanced Mathematics, vol. 19 (American Mathematical Society, Providence/International Press, Boston, 2001). https://doi.org/10.1090/amsip/019
16. E.J. Straube, *Lectures on the $\mathcal{L}^2$-Sobolev Theory of the $\overline{\partial}$-Neumann Problem*. ESI Lectures in Mathematics and Physics (European Mathematical Society (EMS), Zürich, 2010). https://doi.org/10.4171/076
17. T. Ohsawa, *L^2 Approaches in Several Complex Variables*. Springer Monographs in Mathematics, 2nd edn. (Springer, Tokyo, 2018). https://doi.org/10.1007/978-4-431-56852-0
18. J.-P. Demailly, *Complex Analytic and Differential Geometry* (2012). Open Access Online https://www-fourier.ujf-grenoble.fr/~demailly/manuscripts/agbook.pdf

Scattering on Periodic Lattices

2

Maxwell T. Hansen

Abstract

Lattice regularization of quantum field theories (lattice QFT) is a powerful tool for understanding such theories when analytic methods are not applicable. However, the utility of numerical results can be affected by two issues: (i) calculations are necessarily performed in a finite-volume spacetime, and (ii) imaginary-time correlation functions are estimated. Both aspects play a particularly important role for multiparticle observables, including scattering and decay amplitudes.

In these lectures we will give an overview, together with various detailed examples, on the topic of extracting scattering amplitudes using numerical lattice field theory. We will focus on the strategy of using the finite volume as a tool rather than an unwanted artifact, and of applying generic field-theoretic relations between finite-volume quantities and infinite-volume amplitudes. Specific topics include the challenges of two-to-three and three-to-three amplitudes and the role of anomalous thresholds. The lectures will include an overview of how the relations are derived and their limitations, as well as a small (and biased) selection of numerical results.

2.1 Introduction and Exponentially Suppressed Finite-Volume Effects

Maxwell T. Hansen

We have overwhelming evidence for quantum chromodynamics (QCD) describing strong interactions over a broad range of energy scales. A recipe for strong-force

M. T. Hansen (✉)
School of Physics and Astronomy, University of Edinburgh, Edinburgh, UK
e-mail: maxwell.hansen@ed.ac.uk

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

N. Arkani-Hamed et al. (eds.), *Records from the S-Matrix Marathon*,
Lecture Notes in Physics 1041, https://doi.org/10.1007/978-3-031-90352-6_2

predictions can be summarized as follows: We start with the Lagrangian defining QCD,

$$\mathcal{L}_{\text{QCD}} = \sum_f \bar{\Psi}_f (i \not{D} - m_f) \Psi_f - \frac{1}{4} G^a_{\mu\nu} G_a^{\mu\nu} , \tag{2.1.1}$$

where f runs over the quark flavors, Ψ_f are the quark fields, m_f are the quark masses, $G^a_{\mu\nu}$ are the gluon field strengths, and $\not{D} = D_\mu \gamma^\mu$ where D_μ is the covariant derivative. We then add to this a specific calculational machinery, such as perturbation theory, effective field theory, or in the focus of this work: numerical lattice QFT. The method of lattice QFT, which when applied to QCD is known as lattice QCD, is the only known systematically improvable non-perturbative method for quantum field theories that admit no analytic solution. In practice this means that in the low energy regime, where quarks and gluons are no longer the relevant degrees of freedom, it is the only available tool for making precise and reliable predictions. As a final step in the setup, we must sacrifice a small set of experimental inputs, such as the masses of the pion, kaon and the omega baryon, in order to set the free parameters of the calculation.

Having made these choices we can then, in principle, make a wide range of precision pre- or post-dictions, for example for the masses of other hadrons (the name collectively given to the low-energy degrees of freedom, bound states of quarks and anti-quarks), hadronic decay constants and form factors, the anomalous magnetic moment of the muon, and the value of quark masses and the strong coupling constant α_s defined in a given renormalization scheme at a given scale.

By now the field of lattice QCD is very advanced. This is well-captured in the reports of the Flavor Lattice Averaging Group (FLAG) [1] which summarize and average the most reliable calculations, providing precise and reliable values of many relevant experimentally relevant parameters. In addition to providing overwhelming evidence for QCD as the theory of the strong force, lattice QCD can provide a tool for new physics searches in the form of Standard Model precision tests.

Taking a step back, lattice QFT is a non-perturbative regularization of a QFT, using a definition that is well-suited to numerical evaluation. A schematic of how it fits in the broader context of QFTs is shown in Fig. 2.1. The key idea is to render the quantum path-integral finite dimensional, and then evaluate it using Monte Carlo importance sampling or potentially more advanced algorithms.

This program comes with are some clear limitations. Schematically, we can write the desired path integral as

$$\text{observable} = \int \mathcal{D}\phi \exp\left[i \int d^4x \mathcal{L}[\phi(x)] \right] \times \left[\text{interpolator for observable} \right] , \tag{2.1.2}$$

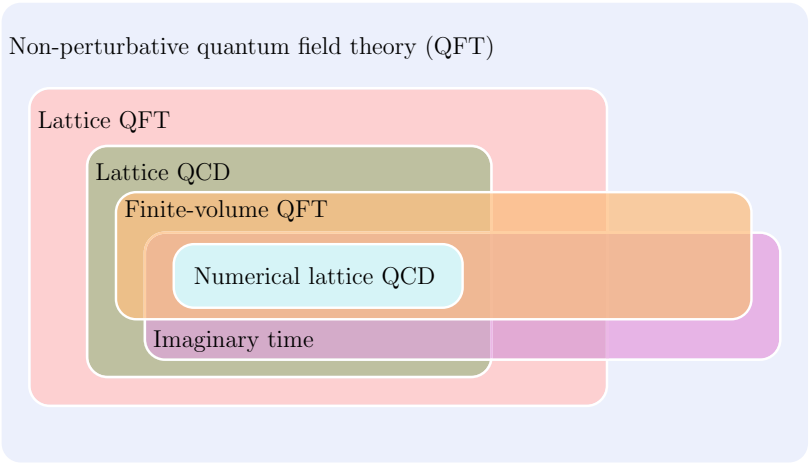
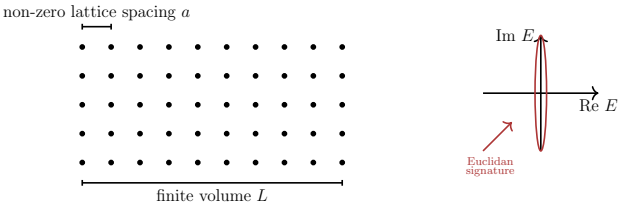


Fig. 2.1 Overview of the interplay of lattice quantum field theory and lattice QCD with finite volume and imaginary time

and the lattice path integral as

$$\text{observable?} = \int d^N\phi \exp\left[-\sum_n^{L/a,T/a} \mathcal{L}_{\text{Euc.}}[\phi(an)]\right] \times \left[\text{interpolator for observable}\right]. \tag{2.1.3}$$

The second line is meant to indicate that we have to make three modifications to our quantum field theory: (i) introduce a non-zero lattice spacing a, (ii) work in a finite spatial volume with length L and a finite temporal extent T, (iii) take time to be imaginary, i.e. we work in Euclidean signature:



(2.1.4)

The last point is important to avoid highly oscillatory integrals, that would make numerical evaluation intractable. It is also common to take some quark masses (set by meson masses like the pion and the kaon) to be higher in lattice simulations to improve the numerics (e.g., $M_{\pi,\text{lattice}} > M_{\pi,\text{our universe}}$), although working at physical mass values is becoming increasingly common.

2.1.1 Processes with QCD-Stable Hadrons

It is useful to identify three types of processes with QCD-stable hadrons:

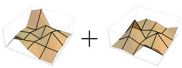
1. Decay constants. These correspond to matrix elements of the form $\langle 0|\mathcal{J}|1\rangle$, where $\langle 0|$ represents the vacuum, $\mathcal{J}$ a local current (in practice constructed via a quark-anti-quark bilinear), and $|1\rangle$ represents a single-hadron state. Examples include the pion (f_π), kaon (f_K) or B-meson (f_B) decay constants.
2. Form factors. These correspond to matrix elements of the form $\langle 1'|\mathcal{J}|1\rangle$. Examples include the form factor between a kaon and pion at a scale q^2: $f_+^{K^0\pi^-}(q^2)$.
3. Mixing parameters, corresponding to matrix elements like $\langle \bar{1}|\mathcal{H}^{\Delta F=2}|1\rangle$, e.g. the mixing parameter of a neutral kaon and its anti-particle. In contrast to the previous case, the key different here is the nature of the local operator $\mathcal{H}^{\Delta F=2}$, which now consists of two quarks and two anti-quarks.

Each of these types of processes and corresponding observables are well under control in modern lattice QCD calculations and the latest FLAG report provides a summary as well as details of methods [1].

Focusing on the decay constants (item number 1), we now give a high level summary of how a lattice QCD calculation proceeds, favoring the basic idea over technical details and targeting someone working on aspects of quantum field theory or the S-matrix from a very different perspective.

The first step is to identify a Euclidean correlation function that will give access to the target observable. In the case of f_π this can be achieved by combining the axial vector current, denoted $A_\mu^{\text{bare}}(0)$ (where bare indicates that renormalization is needed), with an operator $\pi_{\mathbf{p}}(-\tau)$, that has the quantum numbers of a pion with a given momentum $\mathbf{p}$ and a given value of Euclidean time $-\tau$.

One then uses the path-integral formalism to numerically evaluate the QCD vacuum expectation value of $A_\mu^{\text{bare}}(0)\pi_{\mathbf{p}}(-\tau)$, subject to the modifications mentioned above. This numerical sampling can be represented by a cartoon as follows:

$$\langle A_\mu^{\text{bare}}(0)\pi_{\mathbf{p}}(-\tau)\rangle_{T,L,m_q,a} =$$  $$+\cdots$$ (2.1.5)

The right-hand side reflects the fact that we have a finite set of gauge field configurations from Monte Carlo importance sampling and evaluating $A_\mu^{\text{bare}}(0)\pi_{\mathbf{p}}(-\tau)$ on each and summing gives the expectation value. The subscripts on the left-hand side remind us that the result depends on the quark masses (m_q), the spacetime volume, and the lattice spacing.

The bare current requires renormalization. We will not concern ourselves with the details but only note this is understood. We can thus relate our correlator estimate to the physical decay constant f_π by the relation

$$Z^{\text{renorm}}\langle A_\mu^{\text{bare}}(0)\pi_{\mathbf{p}}(-\tau)\rangle_{T,L,m_q,a} = \langle A_\mu^{\text{renorm}}(0)\mathrm{e}^{-\hat{H}\tau}\pi_{\mathbf{p}}(0)\rangle\,, \tag{2.1.6}$$

$$= \sum_n \langle A_\mu^{\text{renorm}}(0)|n\rangle \mathrm{e}^{-E_n\tau} \langle n|\pi_{\mathbf{p}}(0)\rangle \tag{2.1.7}$$

$$\xrightarrow[T\gg\tau\gg\delta E_\pi]{} Z_\pi \mathrm{e}^{-E_\pi\tau}\, ip_\mu f_\pi(T, L, m_q, a)\,. \tag{2.1.8}$$

On the second line we have inserted a complete set of states with the relevant quantum numbers. (This really is a discrete sum over gapped states due to the finite volume L.) The third line is a statement of the asymptotic behavior of the correlation function at large Euclidean time. The exponential suppression of excited states is a key feature of the Euclidean time formalism. The decay constant f_π is then extracted by fitting the large-time behavior of the correlation function to the expected form. It is essential to emphasize that, in this case the Euclidean time is not a disadvantage and also that the matrix element defining f_π has no memory of the metric signature we started with.

The next important observation is that the physical decay constant is only recovered from $f_\pi(T, L, m_q, a)$ by removing all unwanted effects of the modifications indicated by the four arguments: finite temporal and spatial extent, non-zero lattice spacing, and unphysical quark masses. The procedure is well understood and can be summarized as:

$$\lim_{T,L\to\infty} \lim_{a\to 0} f_\pi(T, L, m_q^{\text{phys}}, a) = f_\pi^{\text{phys}} \tag{2.1.9}$$

If we started with unphysical quark masses, we have here assumed that we can also extrapolate the decay constants to the physical quark masses. A summary of decay constants for D and B mesons is shown in Fig. 2.2.

Audience Question 2.1.1. How can you take the infinite time and volume limit?

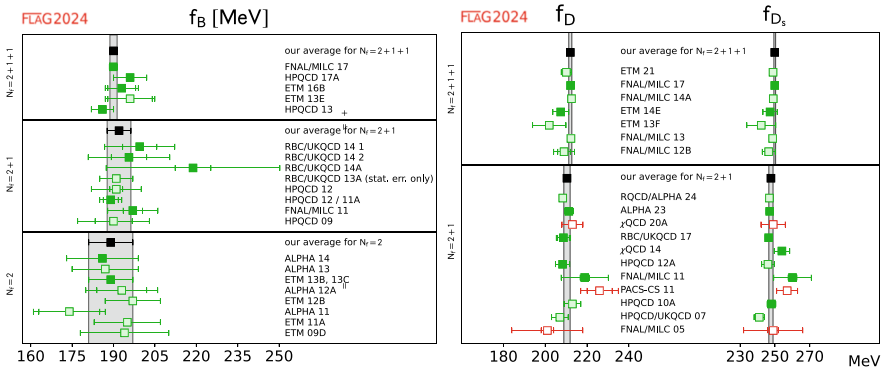


Fig. 2.2 A summary of decay constants. Reprinted with permission from [1]. © 2024, The Author(s). All rights reserved. Data from [2–18]

Answer: To take the limit, one can fit known large L and large T expansions to the data. Such expansions are generally of the form $O(L,T) = O(\infty,\infty) + \Delta O(L,T)$ where $O(\infty,\infty)$ is a fit parameter and $\Delta O(L,T)$ depends on additional parameters. So the extrapolation to $T, L \to \infty$ really amounts to taking the value of $O(\infty,\infty)$, provided one can get a reliable fit. There is a lot of formal work in understanding if there could be surprises beyond the sampled data, both for the $L, T \to \infty$ limit and the $a \to 0$ limit.

2.1.2 Exponentially Suppressed Finite-Volume Effects

Motivated by the preceding sketch of a lattice QCD calculation and by the question concerning $L, T \to \infty$, we now turn to the topic of finite-volume effects. This is important as an analytic description of the T, L dependence is needed to estimate the limit from numerical data. More importantly, this will set the stage for the following sections, in which finite-L effects become a tool rather than an unwanted artifact in order to extract scattering amplitudes from lattice QCD calculations.

We begin with a stable particle, like the pion in QCD. The setup here is a continuous (not discretized $\to$ no lattice) spacetime compactified to $\mathbb{T}_3 \times \mathbb{R}$, where we assume periodicity in L leading to the torus $\mathbb{T}_3$. In this setup the spatial momentum components are restricted to take values that are integer three-vector multiples of $\frac{2\pi}{L}$, $\mathbf{p} \in \frac{2\pi}{L}\mathbb{Z}^3$. The time direction is taken to be infinite and the lattice spacing set to zero as we assume these effects are either small or have been removed from the numerical data before the results derived here are applied.[1]

As a first example we consider finite-volume effects on a spin-zero particle in next-to-leading-order $\lambda\phi^4$ theory. The relevant relation is

$$M(L)^2 = M^2 + \quad + \mathcal{O}(\lambda^2), \tag{2.1.10}$$

where $M(L)$ is the mass of the particle in a finite volume, M is the bare mass and the one-loop correction is given by

$$= \frac{-i\lambda}{2}\frac{1}{i}\int\frac{dk^0}{2\pi}\frac{1}{L^3}\sum_{\mathbf{k}}\frac{1}{-(k^0)^2+\mathbf{k}^2+M^2-i\varepsilon}\,, \tag{2.1.11}$$

$$= \frac{\lambda}{2}\int\frac{dk_4}{2\pi}\frac{1}{L^3}\sum_{\mathbf{k}}\frac{1}{k^2+M^2}\,, \tag{2.1.12}$$

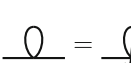
where in the second line we have Wick rotated to Euclidean space without changing the value of the integral.

[1] More precisely, the Monte Carlo importance sampling implies that data generated in lattice QCD calculations have statistical uncertainties and the data can be viewed as continuum, infinite-T, provided the effects of the lattice spacing and temporal extent are smaller than the statistical uncertainties.

We can next introduce a Schwinger parameter α and apply the Poisson summation formula to write

$$= \frac{\lambda}{2} \sum_{\mathbf{n}} \int \frac{\mathrm{d}^4 k}{(2\pi)^4} \mathrm{e}^{iL\mathbf{n}\cdot\mathbf{k}} \int_0^\infty \mathrm{d}\alpha \, \mathrm{e}^{-\alpha(k^2+M^2)} \,, \tag{2.1.13}$$

where the sum runs over all integer three-vectors $\mathbf{n}$. (Note that these are not in direct correspondence with the momentum components $\mathbf{k}$ as they are instead the Poisson modes.) We can next evaluate the integral over k and then split the expression into its $L \to \infty$ limit and a remainder:



$$= \quad + M^2 \frac{\lambda}{32\pi^2} \sum_{\mathbf{n}\neq 0} \int_0^\infty \frac{1}{\alpha^2} \mathrm{e}^{-\alpha - \frac{1}{\alpha}[M^2 L^2 \mathbf{n}^2/4]} \,. \tag{2.1.14}$$

Note that for non-zero $\mathbf{n}$ the integral is convergent. By contrast the $\mathbf{n} = 0$ term is divergent near $\alpha = 0$, corresponding to the momentum space ultraviolet divergence of the original expression. Regardless of how this divergence is regulated, when we combine the integrated loop with the bare mass, we recover the physical (pole) mass to the order we are working. Thus we reach

$$\frac{M(L)^2 - M_{\text{phys}}^2}{M_{\text{phys}}^2} = \frac{\lambda}{32\pi^2} \sum_{\mathbf{n}\neq 0} \int_0^\infty \frac{1}{\alpha^2} \mathrm{e}^{-\alpha - \frac{1}{\alpha}[M^2 L^2 \mathbf{n}^2/4]} + \mathcal{O}(\lambda^2) \,, \tag{2.1.15}$$

where we have also used the M^2 multiplying the integral can be replaced by M_{phys}^2 to the order we are working.

The final step is to evaluate the integral

$$\frac{M(L)^2 - M_{\text{phys}}^2}{M_{\text{phys}}^2} = \frac{\lambda}{32\pi^2} \sum_{\mathbf{n}\neq 0} \frac{K_1(ML|\mathbf{n}|)}{ML|\mathbf{n}|} + \mathcal{O}(\lambda^2) \,, \tag{2.1.16}$$

where K_1 is a modified Bessel function of the second kind. The asymptotic behavior of this function is $K_1(x) \sim \sqrt{\frac{\pi}{2x}} \mathrm{e}^{-x}$ for $x \gg 1$. This gives the exponential suppression of finite-volume effects and in particular

$$\frac{M(L)^2 - M_{\text{phys}}^2}{M_{\text{phys}}^2} = \frac{6\lambda}{32\pi^2} \frac{1}{ML} \sqrt{\frac{\pi}{2}} \frac{1}{\sqrt{ML}} \mathrm{e}^{-ML} + \mathcal{O}(\lambda^2, \mathrm{e}^{-\sqrt{2}ML}, \mathrm{e}^{-ML}/L^{5/2}) \,. \tag{2.1.17}$$

This concludes our first example of finite-volume effects in a quantum field theory. We identify two key take-home messages of the result: (i) the effects are calculable order by order in perturbation theory, and are exponentially suppressed in the

volume, and (ii) the coefficient of the leading exponential depends on the coupling constant λ. This second point is the first hint that finite-volume effects can be used as a tool to extract information about the underlying theory, in particular quantities related to scattering. In fact if $M(L)^2$ is a well defined and non-perturbative quantity, the same must be true for the coefficient of the leading exponential. So as we work to higher orders this will not be some scheme dependent coupling constant, but a more universal observable. This was worked out explicitly by Lüscher in [19] in the context of masses. We will now turn to some related results for the anomalous magnetic moment of the muon.

At this stage it is useful to identify three categories of results: fixed-order perturbative results, all-orders perturbative results, and non-perturbative results. The above example falls into the first category and most of the work presented in these lectures falls in the second. But it is possible to give one example of an almost completely non-perturbative derivation, for the anomalous magnetic moment of the muon

$$a_\mu = \frac{(g-2)_\mu}{2} \,. \tag{2.1.18}$$

where the g-factor, g_μ is a number that describes how the muon's magnetic moment relates to its spin, and should be close to 2 as predicted by Dirac's theory for spin-half leptons.

This is based on work with Agostino Patella [20, 21] taking great inspiration from [19].

At the time of the lectures, the experimental value and Standard Model prediction were given by Aoyama et al. [22]

$$a_\mu^{\text{experiment}} = 0.001\,165\,920\,59(22)\,, \qquad \text{FNAL 2023, World average} \tag{2.1.19}$$

$$a_\mu^{\text{Standard model}} = 0.001\,165\,918\,10(43)\,, \qquad \text{2020 White paper} \tag{2.1.20}$$

The theory uncertainty is dominated by the uncertainty from a QCD contribution called the hadronic vacuum polarization (HVP). This quantity can also be computed in lattice QCD, via [23]

$$a_\mu^{\text{HVP,LO}}(T,L) = \frac{2\alpha^2}{m_\mu^2} \int_0^{T/2} dx_4\, \widehat{\mathcal{K}}(m_\mu x_4) G_{T,L}(x_4)\,, \tag{2.1.21}$$

where m_μ is the muon mass, $\alpha \approx 1/137$ is the fine-structure constant and $\widehat{\mathcal{K}}$ is a continuous kernel function. The function $G_{T,L}(x_4)$ is the Euclidean time-ordered electromagnetic current-current correlator in a finite volume L and temporal extent T.

We can write the correlator as [20, 21]

$$G_{T,L}(x_4) = -\frac{1}{3}\int_{L^3} \mathrm{d}^3\mathbf{x}\langle j_k(x_4,\mathbf{x})j_k(0)\rangle_{T,L}\,, \tag{2.1.22}$$

$$= -\frac{1}{3}\int_{L_\perp} \mathrm{d}^2\mathbf{x}_\perp \int_0^L dx_3 \frac{\mathrm{Tr}\left[\mathrm{e}^{-(L-x_3)\hat{H}} j_\mu(\tilde{x})\mathrm{e}^{-x_3\hat{H}} j_\mu(0)\right]}{\mathrm{Tr}\left[\mathrm{e}^{-L\hat{H}}\right]}\,. \tag{2.1.23}$$

The second line here already represents the essential non-perturbative idea. By taking advantage of the equivalence of the four Euclidean directions, we can quantize along the 3 direction with a hamiltonian $\hat{H}$ whose eigenstates satisfy the $L \times L \times T$ boundary conditions.

The next step is to extend to a more general finite-volume, represented by the matrix $\mathbb{L} = \mathrm{diag}(L_1, L_2, L_3, L_4)$ and then take the limit $L_1, L_2, L_4 \to \infty$ to obtain $G_{3,L_3}(x_4)$. The finite-volume effects of this single compactified direction are then given by $\Delta G_{3,L_3}(x_4) = G_{3,L_3}(x_4) - G_{3,\infty}(x_4)$. In addition, one can apply a cheat by asserting the plausible relation that

$$\Delta G_L(x_4) = 3\Delta G_{3,L}(x_4) + \mathcal{O}(\mathrm{e}^{-\sqrt{2}M_\pi L})\,, \tag{2.1.24}$$

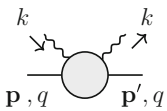
where $\Delta G_L(x_4) = \lim_{T\to\infty} G_{T,L}(x_4) - \lim_{T,L\to\infty} G_{T,L}(x_4)$. In words, this relation asserts that the finite-volume effects in the full L^3 volume are equal, at leading order, to three times the finite-volume effects of a theory with a single compactified direction of length L. This can be proven to all orders in a generic perturbative expansion but we are not aware of a non-perturbative proof.

Carefully applying the $L_1, L_2, L_4 \to \infty$ limits in the above expression (see [20, 21] for details), we can then write

$$\Delta G_3(x_4|L) = -\frac{1}{3}\sum_{q=0,\pm1}\int \frac{\mathrm{d}p_3}{2\pi}\frac{\mathrm{e}^{-L\sqrt{M_\pi^2+p_3^2}}}{4\pi L}$$
$$\times \int \frac{dk_3}{2\pi}\cos(k_3x_4)\,\mathrm{Re}\,T_q(-k_3^2, -k_3p_3)\,. \tag{2.1.25}$$

The function T_q is the forward Compton amplitude $\pi\gamma \to \pi\gamma$ for a pion with charge q,

$$T_q(k^2, k\cdot p) = \lim_{\mathbf{p}'\to\mathbf{p}} \tag{2.1.26}$$



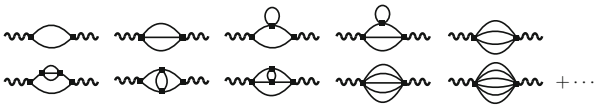
This is now the second example of how finite-volume effects are dictated by scattering amplitudes.

This relation, as well as higher order corrections, can alternatively be derived by switching now to the all orders diagrammatic expansion of the correlator or of $a_\mu^{\text{HVP,LO}}(T, L)$. For asymptotically large T and L, we expect the correlator to have the following general structure:

$$\begin{aligned} a_\mu^{\text{HVP,LO}}(T,L) - a_\mu^{\text{HVP,LO}} = {} & \mathcal{O}(\mathrm{e}^{-M_\pi L}) + \mathcal{O}(\mathrm{e}^{-\sqrt{2}M_\pi L}) + \mathcal{O}(\mathrm{e}^{-\sqrt{3}M_\pi L}) \\ & + \mathcal{O}(\mathrm{e}^{-\sqrt{2+\sqrt{3}}M_\pi L}) + \dots \\ & + \mathcal{O}(\mathrm{e}^{-M_\pi T}) + \mathcal{O}(\mathrm{e}^{-\frac{3}{2}M_\pi T}) + \dots \\ & + \mathcal{O}(\mathrm{e}^{-M_\pi \sqrt{T^2+L^2}}) + \dots \\ & + \mathcal{O}(\mathrm{e}^{-M_K L}) + \dots \end{aligned} \tag{2.1.27}$$

The leading-order term, i.e. $\mathcal{O}(\mathrm{e}^{-M_\pi L})$ was computed in [20]. See Table 2.1. The next subleading ones at $\mathcal{O}(\mathrm{e}^{-\sqrt{2}M_\pi L})$, $\mathcal{O}(\mathrm{e}^{-\sqrt{3}M_\pi L})$ as well as $\mathcal{O}(\mathrm{e}^{-M_\pi T})$ were computed in [21].

The all-orders diagrammatic expansion is represented schematically by

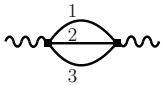
$a_\mu(L) =$  (2.1.28)

where the external wavy lines represent the two-vector currents defining $G_{T,L}(x_4)$ and the internal lines are the pion propagators. The key idea is that we think in terms of a generic low-energy effective field theory, where the pion is the lightest degree of freedom. We make no reference to the details of the theory, e.g. the values of couplings, but simply use the diagrams to keep track of singularities that dictate the leading L and T dependence.

Focusing here on finite-L, note as above that the loop momenta are summed and this can be written using the Poisson summation formula

$$\frac{1}{L^3}\sum_{\mathbf{k}} \to \sum_{\mathbf{n}} \int \frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^3} \mathrm{e}^{iL\mathbf{n}\cdot\mathbf{k}}\,. \tag{2.1.29}$$

The vector $\mathbf{n} = (n_x, n_y, n_z)$ can be interpreted as counting how pions wrap the torus. There are some subtleties in assigning the Poisson modes, for example,



$$\begin{matrix} \mathbf{n}_1 = (0,0,1) \\ \mathbf{n}_2 = (0,0,1) \end{matrix} \quad \equiv \quad \begin{matrix} \mathbf{n}_2 = (0,0,0) \\ \mathbf{n}_3 = (0,0,1) \end{matrix}\,, \tag{2.1.30}$$

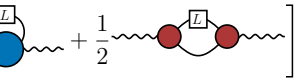
Table 2.1 For a set of model assumptions of the value of $a_\mu^{\mathrm{HVP,LO}}$ and the form of the Compton amplitude, this table shows the numerical estimates for the finite-L dependence. Note that the $\mathrm{e}^{-\alpha_n M_\pi L}$-series converges slowly. The last column estimates the percentage of the finite-volume effects. The goal is to get sub 1% precision for theoretical relevance. Reproduced under CC-BY-4.0 license from [20]. © 2019, The Author(s)

$-100\times\Delta a_{\mathrm{s}}(L)/a_\mu^{\mathrm{HVP,LO}}$									
$M_\pi L$	$\lvert\mathbf{n}\rvert=1$	$\sqrt{2}$	$\sqrt{3}$	2	$\sqrt{5}$	$\sqrt{6}$	$2\sqrt{2}$	3	$\sum_{\mathbf{n}}$
4	1.26	1.16	0.317	0.104	0.193	0.0944	0.0128	0.0174	3.17
5	0.852	0.428	0.0813	0.0199	0.0287	0.0112	0.00102	0.00117	1.42
6	0.461	0.141	0.0189	0.00349	0.00394	0.00124	0.0000764	0.0000735	0.630
7	0.226	0.0433	0.00417	0.000582	0.000515	0.000130	5.46×10^{-6}	4.41×10^{-6}	0.274
8	0.104	0.0128	0.000883	0.0000936	0.0000652	0.0000132	3.79×10^{-7}	2.57×10^{-7}	0.118

are equivalent and should not be counted separately. This can be understood as a gauge redundancy, and a gauge fixing allows one to correctly catalog effects [19,21].

The leading contributions come from only one pion wrapping the torus,

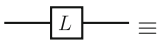
$$\Delta a_\mu(L) = \int_0^\infty \mathrm{d}x_0\, \mathcal{K}(x_0) \Bigg[\quad + \frac{1}{2} \quad \Bigg] + \mathcal{O}(\mathrm{e}^{-1.9 M_\pi L}) \,, \tag{2.1.31}$$



where the shaded blobs represent one-particle irreducible vertices, defined by their external legs. The neglected expontial has a coefficient of $\sqrt{2+\sqrt{3}} \approx 1.9$ associated with the leading effect of two-pions wrapping the torus [21].

All the L dependence is inside the modified propagator

$$\equiv \sum_{\mathbf{n}} \frac{\mathrm{e}^{iL\mathbf{n}\cdot\mathbf{p}}}{p^2 + M_\pi^2 + \Sigma(p^2)} \,. \tag{2.1.32}$$



Naively it looks like there might be an oscillatory L dependence but this integrates to exponential decay as can be seem by deforning the integration contour. The dominant scaling comes from the nearest singularity which is at $p^2 = -M_\pi^2$, along the imaginary axis for a Euclidean theory. The countour is defored as

(2.1.33)

and the integral is dominated by the pole at $p^2 = -M_\pi^2$. This gives another perspective on why leading L dependence is dictated by scattering amplitudes. The key points in this context are (i) the leading finite-volume effects can be expressed in terms of irreducible vertex functions and an L-dependent modified pion propagator and (ii) the leading L dependence is dictated by the nearest singularity in the modified propagator, and identifying this contribution sets the L-dependent pion on its mass shell.

The result is that the sum of all single winding terms give

$$\Delta a_\mu(L) = -\frac{\alpha^2}{m_\mu^2} \sum_{\mathbf{n}\neq\mathbf{0}} \int \frac{dp_3}{2\pi} \frac{\mathrm{e}^{-|\mathbf{n}|L\sqrt{M_\pi^2+p_3^2}}}{2\pi L|\mathbf{n}|} \int_0^\infty dx_0\, \widehat{\mathcal{K}}(m_\mu x_0) \tag{2.1.34}$$

$$\times \int \frac{dk_3}{2\pi} \cos(x_0 k_3) \mathrm{Re}\, T(-k_3^2, -k_3 p_3) + \mathcal{O}(\mathrm{e}^{-\sqrt{2} M_\pi L}) \,, \tag{2.1.35}$$

where T is the Compton amplitude defined above, summed over the pion charges.

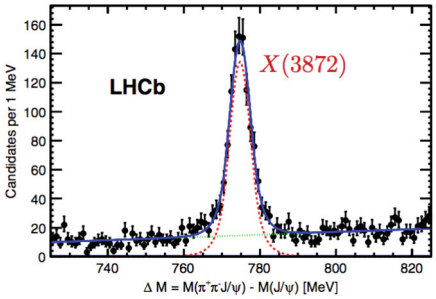


Fig. 2.3 Plot of the event count as a function of mass difference ΔM in $B^+ \to J/\psi K^+\pi^+\pi^-$ decays. Black points represent data from LHCb. Total, signal, and background fits for the $X(3872)$ signal are plotted in blue, red, and green respectively. Reprinted under CC-BY-3.0 license from [25]. © 2015, The Author(s)

2.2 Two-Particle Scattering

Maxwell T. Hansen

In the previous section, we have seen examples of single-hadron observables that can be reliably calculated to %-level precision using lattice QCD. We have focused on how the effect of the finite volume can be quantified and removed and have also emphasized that the leading L-dependence is related to scattering amplitudes. This gives a hint of a more general structure, to be explored in this section in the context of two-to-two scattering, with a focus on scattering low-energy degrees of freedom in QCD, i.e. hadrons.

2.2.1 Multi-Hadron Observables

We begin by motivating the study of multi-hadron observables more generally. One reason why we might care about such observables is that they are relevant for understanding the plethora of largely unexpected bumps seen recently in experimental cross-section data. At least some of these are expected to correspond to exotic QCD resonances, collectively referred to as XYZ states, and some are also tetra- and penta-quark candidates.[2] As an example, Fig. 2.3 shows an exotic meson candidate, $X(3872)$, in experimental data from the LHCb collaboration [25]. Detection of such states was not entirely expected, but it should not be viewed as a major surprise. In short, we understand QCD too poorly in this regime to have any real basis for expectations.

Multi-hadron observables are additionally relevant for more completely understanding the weak interactions. For example, by analyzing $D \to \pi\pi$ and $K\bar{K}$ decays, LHCb recently reported CP asymmetry through a parameter called ΔA_{CP}

[2] See e.g. [24] for a recent review.

with the value $\Delta A_{CP} = (-15.4 \pm 2.9) \times 10^{-4}$ resolved to be significantly different from zero for the first time in this channel [26]. Such discoveries could have connections to new physics beyond the Standard Model, e.g. in the context of explaining matter-antimatter asymmetry. However, we cannot say anything concrete until we have a reliable Standard Model prediction for ΔA_{CP}. This issue is also related to resonances. For example, it has been suggested that a nearby resonance called the $f_0(1710)$ could provide a significant Standard Model enhancement to ΔA_{CP} [27]. To make a quantitative statement, we need a method that treats this messy physics rigorously from first principles.

We have stated that peaks in cross-sections correspond to resonances, but without defining what resonances are. It is easy to reach misconceptions, for example when we say "states" a notation like this comes to mind:

$$|X\rangle, \quad |\rho\rangle, \quad |f_0\rangle\,. \tag{2.2.1}$$

This is imprecise as these are not really defined in the QCD Fock space.

Instead the meaningful "states" are asymptotic multi-hadron states. At low energies, QCD consists of hadronic degrees of freedom such as

$$\pi \sim \bar{u}d, \quad K \sim \bar{s}u, \quad p \sim uud\,. \tag{2.2.2}$$

A two-pion state, for example, can then be defined by asymptotically seperating two pions with definite momenta at early times. The construction is connected to the Lehmann–Symanzik–Zimmermann reduction formula or more rigorously to Haag–Ruelle scattering theory. The S-matrix is then defined as an overlap between multi-hadron asymptotic states.

In general the S-matrix depends on the center-of-mass-frame (CMF) energy E_{cm}, angular variables, and (in the case of more than two incoming or outgoing particles) on invariant masses of subsystems. But if we focus on two-to-two scattering, e.g. $\pi\pi \to \pi\pi$, the degrees of freedom reduce to the CMF energy and a single angular variable, that can be used to define angular momentum. Moreover, the S-matrix is diagonal in the angular momentum space and is unitary. Thus, for $E_{\text{cm}} < 4M_\pi$, the S-matrix is just populated with phases:

$$S(s) \equiv \langle \pi\pi, \text{out}| \quad \begin{array}{c|ccc} & \multicolumn{3}{c}{|\pi\pi, \text{in}\rangle} \\ & \text{} & \text{  } & \text{  } \\ \hline \text{  } & e^{2i\delta_0(s)} & 0 & 0 \\ \text{  } & 0 & e^{2i\delta_1(s)} & 0 \\ \text{  } & 0 & 0 & e^{2i\delta_2(s)} \end{array} \tag{2.2.3}$$

where we have introduced the Mandelstam variable $s = E_{\text{cm}}^2$.

There is a distinction between the S-matrix and scattering amplitudes, where we subtract the identity corresponding to disconnected contributions:

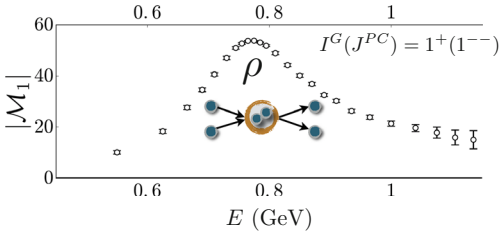
$$\mathcal{M}_\ell(s) \propto \mathrm{e}^{2i\delta_\ell(s)} - 1\,. \tag{2.2.4}$$

We note that, in general the QCD S-matrix represents an enormous space of information. As we raise the energy the matrix will incorporate overlaps of increasingly complicated multi-hadron states, e.g. $|\pi\pi\pi\pi, \text{in}\rangle$, $|K\bar{K}, \text{in}\rangle$, etc.

Let us now come back to QCD resonances. Roughly speaking, they are detected as a bump in the scattering rate which is proportional to the magnitude of the amplitude squared:

$$|\mathcal{M}_\ell(s)|^2 \propto |\mathrm{e}^{2i\delta_\ell(s)} - 1|^2 \propto \sin^2\delta_\ell(s)\,. \tag{2.2.5}$$

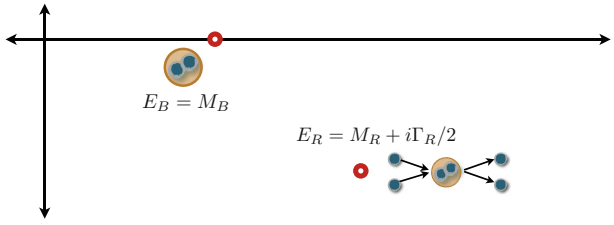
As an illustration, here is a plot for the $\pi\pi \to \pi\pi$ with angular-momentum and parity satisfying $J^P = 1^-$, exhibiting a peak associated with the ρ resonance [28]:



(2.2.6)

The maximum value reached here corresponds to the phase shift $\delta_\ell = \pi/2$ maximising the $\sin^2\delta_\ell$ function.

We reach our final definition of a resonance by considering $|\mathcal{M}_\ell|^2$ in the complex plane. This is analytic in regions of the complex plane so that a continuation can be defined. A resonance is then a pole of the amplitude, represented schematically as:



(2.2.7)

Note that the difference between a bound state and a resonance is that the former is on the real axis, but the latter has a non-zero imaginary part. Our aim will be to understand the amplitude on the real axis and rigorously continue it into the complex plane.

Audience Question 2.2.1. Can you treat scattering of pions with QED corrections where pions are not stable?
Answer: So I suppose you mean neutral pions that decay into two photons. (Note that we cannot form a ρ resonance only from neutral pions.) If we wanted to be particularly pedantic then we could study this in a QCD+QED theory by constructing the four-photon scattering amplitude ($\gamma\gamma\gamma\gamma \to \gamma\gamma\gamma\gamma$). We would then find that the amplitude contains neutral pion poles as a function of both incoming and outgoing two-photon invariant masses. These have a small imaginary part but we can still continue to the pole and perform an amputation analogous to the Lehmann-Symanzik-Zimmermann reduction formula. This would give us the pion-pion scattering amplitude that fully includes the QED corrections and the non-stable nature of the pions. More practically we take advantage of the very different time scales and we can treat pions as stable to a very good approximation in studying the QCD scattering amplitudes.

2.2.2 Bethe–Salpeter Equation and $\mathcal{K}$-Matrix

When it comes to analyzing analyticity properties, it is standard to think about the amplitude $\mathcal{M}_\ell(s)$ instead of the absolute value. In other words, we will analytically continue the amplitude itself. In the pion case, it can have two-, four-, six-particle thresholds and so on. We will confine ourselves to the energies such that we are above the two-particle production but below four:

$$(2m)^2 < s < (4m)^2 \tag{2.2.8}$$

and ask about the analytic structure in this region. Here we use m in place of M_π for the mass, to emphasize that these are general features of two-particle scattering.

There are different ways to think about this. First, the optical theorem tells us

$$\rho(s)|\mathcal{M}_\ell(s)|^2 = \text{Im}\ \mathcal{M}_\ell(s)\,, \tag{2.2.9}$$

where

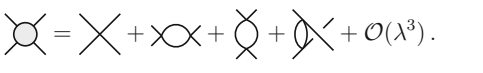
$$\rho(s) = \frac{\sqrt{1-4m^2/s}}{32\pi} \tag{2.2.10}$$

is the two-particle phase space factor. The unique solution to this constraint is

$$\mathcal{M}_\ell(s) = \frac{1}{\mathcal{K}_\ell(s)^{-1} - i\rho(s)}\,. \tag{2.2.11}$$

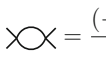
The first factor in the denominator $\mathcal{K}_\ell$ is called the *$\mathcal{K}$-matrix* and it governs the short-distance physics. The second factor $\rho(s)$ is the phase-space branch cut and it tells us about the long-distance behavior. The $\mathcal{K}$-matrix is analytic across $s = 4m^2$, so the only non-analyticity is captured by the $\rho(s)$. This is the key message: the amplitude has a square-root branch cut starting at the two-particle threshold.

Example 2.2.1. As an example, let us consider $\lambda\phi^4$ theory in four dimensions. First, we do this example in infinite volume. The $2 \to 2$ scattering amplitude admits the usual perturbative expansion using Feynman rules:

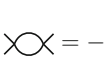
$$i\mathcal{M} =$$  $$= \mathcal{O}(\lambda^3)\,.$$ (2.2.12)

To leading orders, we have the contact diagrams, as well as the s-, t-, and u-channel bubbles.

Let us focus on the s-channel bubble specifically, which is the only one possessing the cut starting at $s = 4m^2$:

$$\text{  } = \frac{(-i\lambda)^2}{2}\left(\frac{1}{i}\right)^2 \int \frac{\mathrm{d}^4k}{(2\pi)^4} \frac{1}{(k^0)^2 - \mathbf{k}^2 - m^2 + i\varepsilon} \frac{1}{(E-k^0)^2 - (\mathbf{P}-\mathbf{k})^2 - m^2 + i\varepsilon}\,,$$ (2.2.13)

where the total incoming energy and momenta are called E and $\mathbf{P}$, respectively. We are going to take a slightly non-standard approach to this diagram, which is to upset the Lorentz invariance. By integrating out the k^0, we are left with a three-dimensional integral:

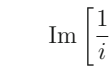
$$\text{  } = -i\frac{\lambda^2}{2} \int \frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^3} \left[\frac{1}{2\omega_k} \frac{1}{E - \omega_k + \omega_{P-k}} \frac{1}{E - \omega_k - \omega_{P-k} + i\varepsilon} + \frac{1}{2\omega_{P-k}} \frac{1}{E + \omega_{P-k} - \omega_k} \frac{1}{E + \omega_{P-k} + \omega_k} \right],$$ (2.2.14)

where

$$\omega_k = \sqrt{\mathbf{k}^2 + m^2}\,.$$ (2.2.15)

This decomposition correspond to the so-called old-fashioned (or time-ordered) perturbation theory. Notice that the $i\varepsilon$ is kept only in the first term because this is the only place where a pole appears on the integration contour.

We are interested in the imaginary part of this diagram (after dividing by the pesky i up front). The only place where the imaginary part appears is precisely in the $i\varepsilon$ term. We can use the famous trick that converts the imaginary part of a propagator into a delta function:

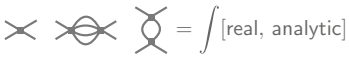
$$\mathrm{Im}\left[\frac{1}{i}\text{  }\right] = -\frac{\lambda^2}{2} \int \frac{\mathrm{d}^3\mathbf{k}_{\mathrm{cm}}}{(2\pi)^3} \frac{1}{4\omega_{k_{\mathrm{cm}}}} (-\pi\delta(E_{\mathrm{cm}} - 2\omega_{k_{\mathrm{cm}}})) = \lambda^2\rho(s)\,.$$ (2.2.16)

We ended up with the integral which is just the two-pion on-shell phase space. This equation is consistent with the general expansion $i\mathcal{M} = i\mathcal{M} + i\mathcal{M}\rho i\mathcal{M} + \ldots$. As we work to higher orders, we find that imaginary parts at a given order are constrained by the real parts at lower orders.

The preceding example is also intended to motivate a skeleton expansion:

$$\mathcal{M}(s) \equiv$$  $$+ \cdots$$ (2.2.17)

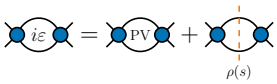
Each solid line represents a fully dressed pion propagator and each blue circle represents a *Bethe–Salpeter kernel* that consists of all diagrams that do not have a two-particle cut, for example:



(2.2.18)

These diagrams have the property of being real and analytic in the elastic region $(2m)^2 < s < (4m)^2$.

This reorganization is useful because only the two-particle loop can contribute a non-zero imaginary part. In a similar fashion to the one-loop example above, one can show that this imaginary part can be expressed in terms of $\rho(s)$. We can isolate it out using the following *cutting rule*:



(2.2.19)

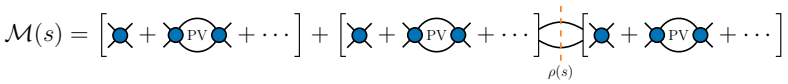
Here, the dashed line correspond to putting the particle on-shell (cutting) and PV denotes the principal value prescription on the propagators.

An essential point is that the imaginary part of the diagram produces a Dirac delta function in the CMF energy, as above and the momenta flowing in the kernels are constrained by the delta function. As a result the legs are set on-shell and the only remaining freedom is the direction of momentum. When this is decomposed in terms of partial waves, then the second term can be written as a product

$$= B(s)_{\ell m,\ell' m'}[i\rho(s)\delta_{\ell'\ell''}\delta_{m'm''}]B(s)_{\ell'' m'',\ell''' m'''}\,,$$ (2.2.20)

where $B(s)$ is the Bethe–Salpeter kernel and repeated indices are summed over.

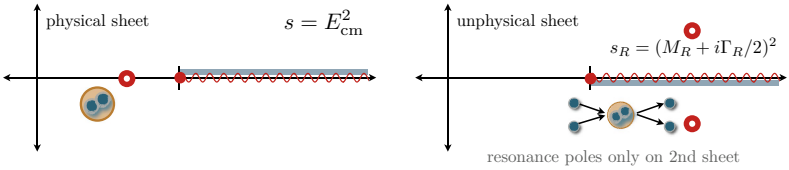
Now we can break up the infinite series (2.2.17) and reorganize it by the number of time the $\rho(s)$ function appears:



$$= \mathcal{K}(s) + \mathcal{K}(s) i\rho(s) \mathcal{K}(s) + \ldots = \frac{1}{\mathcal{K}(s)^{-1} - i\rho(s)} \,. \tag{2.2.21}$$

This sums up to the form we have seen above. Hence, the term in the first bracket *defines* the $\mathcal{K}$-matrix. Our goal in the following section will be to write an analogous formula in the finite-volume context.

Note that the channel opening up at $s = 4m^2$, and in fact every two-particle channel, generates a square-root branch cut. This effectively doubles the number of Riemann sheets. The two sheets are called *physical* (first) and *unphysical* (second). Their analytic structure in the complex s-plane can be represented as:



(2.2.22)

Poles on the first sheet have the interpretation of bound states. Poles on the second correspond to resonances and they are typically displaced into the complex plane as $s_R = (M_R + i\Gamma_R/2)^2$, where M_R and Γ_R are the mass and the width of the resonance respectively. Poles in the upper half-plane on the first sheet would violate causality.

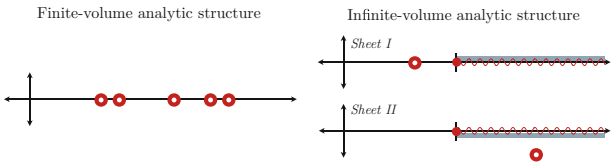
We have already learned that details of analyticity are important for quantitative understanding of the physics involved in the scattering processes. The important lesson we have learned is that it is possible to separate: (i) long-distance kinematic singularities, and (ii) short-distance/microscopic physics that depends on interaction details.

2.2.3 Finite Volume as a Tool

As discussed above, numerical lattice QCD calculations are performed with Euclidean signature (imaginary time), and the resulting correlators do not have an obvious relation to scattering processes that take place in a spacetime with Minkowski signature (real time) [29]. We also cannot use the normal LSZ

procedure, for which we need to be able to access the on-shell kinematics $p_4^2 = -(\mathbf{p}^2 + m^2)$, which is not possible because we only have $p_4 \in \mathbb{R}$.[3]

Another problem is that working with finite volume destroys modifies the analytic properties discussed above. Finite volume discretizes the spectrum, eliminates branch cuts and extra sheets, as well as hides the resonance poles. The picture we obtain is as follows:



(2.2.23)

In a nutshell, the aim now will be to overcome these difficulties and learn how to use the finite volume as a tool for extracting scattering observables. The formalism presented here is based on pioneering work by Lüscher [35,36], together with many subsequent extensions [37–50].

Once again, we are going to use spatial volume with extent L. The momenta are periodic: $\mathbf{p} = \frac{2\pi}{L}\mathbf{n}$ for $\mathbf{n} \in \mathbb{Z}^3$. We assume that L is large enough so that we can neglect $\mathrm{e}^{-M_\pi L}$ corrections. The effects of the time directions will also be negligible.

The scattering of particles at vanishingly small momentum is described by a single number called the scattering length and denoted by a_0. The scattering length appears in two quantities. The first one is the $\ell = 0$ scattering amplitude evaluated on the threshold:

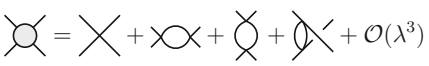
$$\mathcal{M}_{\ell=0}(4m^2) = -32\pi m a_0 \,. \tag{2.2.24}$$

The second quantity is the finite-volume ground state energy [51]

$$E_0(L) = 2m + \frac{4\pi a_0}{mL^3} + \mathcal{O}(1/L^4) \,. \tag{2.2.25}$$

Because of this universality, if we know the latter, we also know the former.

Example 2.2.2. Let us revisit Example 2.2.1, but now including finite-volume effects. We are going to look at the diagrammatic expansion of a function $i\mathcal{M}_L$:



(2.2.26)

It no longer has a direct physical meaning as an amplitude, but it is useful because it contains poles in E at the finite-volume energis. Essentially, we will get the same

[3] For ideas more along the lines of analytic continuation, see for example [30–34].

loop integrand, except now the integral becomes a sum over a discrete set of modes, which means it only has discrete isolated poles. Focusing on the bubble contribution, we get

$$\begin{aligned}
\bigtimes = \frac{(-i\lambda)^2}{2}\left(\frac{1}{i}\right)^2 \int \frac{\mathrm{d}k^0}{2\pi}\frac{1}{L^3}\sum_{\mathbf{k}} \frac{1}{(k^0)^2-\mathbf{k}^2-m^2+i\varepsilon}\frac{1}{(E-k^0)^2-(\mathbf{P}-\mathbf{k})^2-m^2+i\varepsilon} \\
= -i\frac{\lambda^2}{2}\frac{1}{L^3}\sum_{\mathbf{k}}\left[\frac{1}{2\omega_k}\frac{1}{E-\omega_k+\omega_{P-k}}\frac{1}{E-\omega_k-\omega_{P-k}+i\varepsilon}\right. \\
\left. + \frac{1}{2\omega_{P-k}}\frac{1}{E+\omega_{P-k}-\omega_k}\frac{1}{E+\omega_{P-k}+\omega_k}\right].
\end{aligned} \tag{2.2.27}$$

Recall that $\omega_k = \sqrt{\mathbf{k}^2+m^2}$.

To find the ground-state pole we consider what happens as we tune the energy very close to $2m$ plus a small deviation,

$$E = 2m + \delta\,, \tag{2.2.28}$$

where we treat δ as an $\mathcal{O}(\lambda)$ parameter. After plugging it in and expanding in δ, the second term is finite as $\delta \to 0$. However, the first one has a $1/\delta$ pole. We get

$$\bigtimes = -i\frac{1}{L^3}\frac{\lambda^2}{2}\frac{1}{(2M)^2}\frac{1}{\delta} + \mathcal{O}(\delta^0)\,. \tag{2.2.29}$$

This is precisely the same term that before gave us a square-root branch cut. In the finite-volume world it is just a δ pole. We can also write

$$\bigtimes = (i\mathcal{M}_{\mathrm{LO}})if_L(i\mathcal{M}_{\mathrm{LO}}) + \mathcal{O}(\delta^0)\,, \tag{2.2.30}$$

where $\mathcal{M}_{\mathrm{LO}}$ is the tree-level (leading order) contribution and

$$f_L = \frac{1}{8M^2L^3}\frac{1}{\delta} \tag{2.2.31}$$

plays the same role as $\rho(s)$ in infinite volume. The pole means that the perturbative expansion breaks down when δ is a small parameter so we only get a meaningful result by summing such contributions to all orders:

$$\begin{aligned}
i\mathcal{M} &= \bigtimes + \bigtimes + \bigtimes\bigtimes + \bigtimes\bigtimes\bigtimes + \mathcal{O}(\lambda^5) \\
&= \sum_{n=0}^{\infty} i\mathcal{M}_{\mathrm{LO}}(-f_L\mathcal{M}_{\mathrm{LO}})^n = \frac{i}{\mathcal{M}_{\mathrm{LO}}^{-1}+f_L}
\end{aligned} \tag{2.2.32}$$

It looks a bit like the Bethe–Salpeter series we have seen before.

The pole in the resulting expression satisfies

$$f_L = -\mathcal{M}_{\text{LO}}^{-1} \qquad \Longrightarrow \qquad \frac{1}{\lambda} = \frac{1}{8M^2L^3}\frac{1}{\delta}\,. \tag{2.2.33}$$

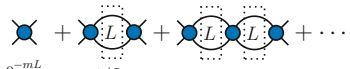
Combined with (2.2.25) and (2.2.28), it allows us to relate the spectrum to the scattering length

$$E_0(L) = 2m + \frac{\lambda}{8m^2L^3} + \mathcal{O}(\lambda^2) = 2m + \frac{4\pi a_0}{mL^3}\left(1 + \mathcal{O}(a_0/L)\right)\,. \tag{2.2.34}$$

In principle, we would have to study this relationship order by order, but one can show that higher orders do not spoil the result above.

2.2.4 General Relation

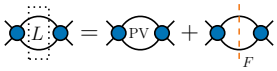
Having in mind the above example, we can proceed to the all-order diagrammatic version of the general result [35–50]. We consider a generic theory with the lightest particle of mass $m \neq 0$. We are going to write down an equation diagrammatically similar to (2.2.17), but physically very different:

$\mathcal{M}_L(P) =$ 

(2.2.35)

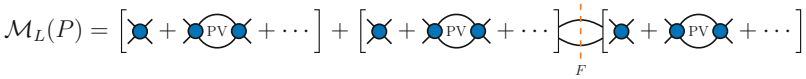
As before, the blue blob denotes the Bethe–Salpeter kernel and solid lines correspond to propagating pions. The dashed box denotes the sum over the discrete momentum modes.

We can now copy what we have done previously in infinite volume. We decompose the on-shell sum into the principal value component and the rest, which we call F:



(2.2.36)

Here, $F(P, L)$ is a matrix of known geometric functions. Hence, once again, we can reorganize the infinite series and define the finite-volume analogue of the $\mathcal{K}$-matrix:

$\mathcal{M}_L(P) =$ 

$$= \frac{1}{\mathcal{K}(s)^{-1} + F(P, L)}\,. \tag{2.2.37}$$

Therefore, the quantization condition becomes

$$\det[\mathcal{K}^{-1}(s) + F(P, L)] = 0\,. \tag{2.2.38}$$

Recall that the Eq. (2.2.38) holds only for two-particle energies, i.e., $s < (4m)^2$, and the matrices are labeled by the angular momenta. There, $\mathcal{K}$ is a diagonal matrix, but F is off-diagonal. It means that in general we don not expect perfect conservation of angular momentum, as the off-diagonal terms will introduce mixing. There are multiple generalizations of (2.2.38) to the cases with non-degenerate masses, multiple channels, spinning particles, etc.

2.2.5 Lattice Results

The only way to make progress is to truncate the $\mathcal{K}$-matrix by shutting off components above some cut-off angular momentum. For example, we can study pion scattering in the single-channel case. The equation we get is

$$\mathcal{K}(s_n)^{-1} = \rho \cot \delta(s_n) = -F(E_n, \mathbf{P}, L)\,, \tag{2.2.39}$$

where F is known. One can then read off the phase shift δ as a function of energy, see Fig. 2.4. It provides a piece of non-perturbative information about QCD: presence of the ρ resonance. Each data point in Fig. 2.4 is obtained from a specific momentum configuration $\mathbf{P}$ and lattice spacing $L = 16, 20, 24$. Different momenta really change the CMF energy at which we probe the system, which is why we get a

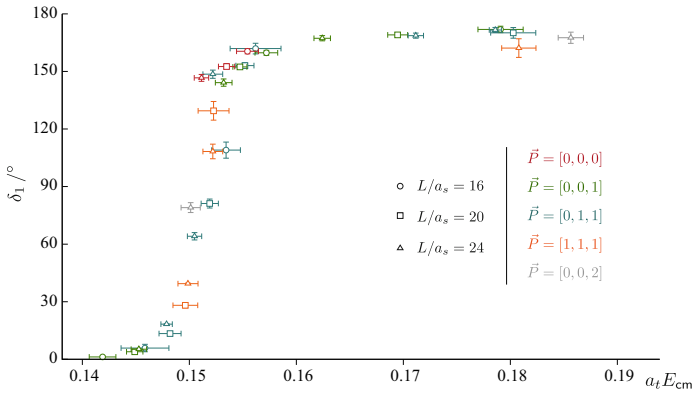


Fig. 2.4 The p-wave elastic phase shift δ_1 (in degrees) for $\pi\pi$ scattering as a function of the CMF energy $E_{\rm cm}$ in lattice units. Each data point is obtained from a different momentum configuration $\mathbf{P}$ and lattice spacing L. The step function in the data is a signal of the ρ resonance. We refer to [52] for more details. Reprinted with permission from [52]. © 2013, American Physical Society. All rights reserved

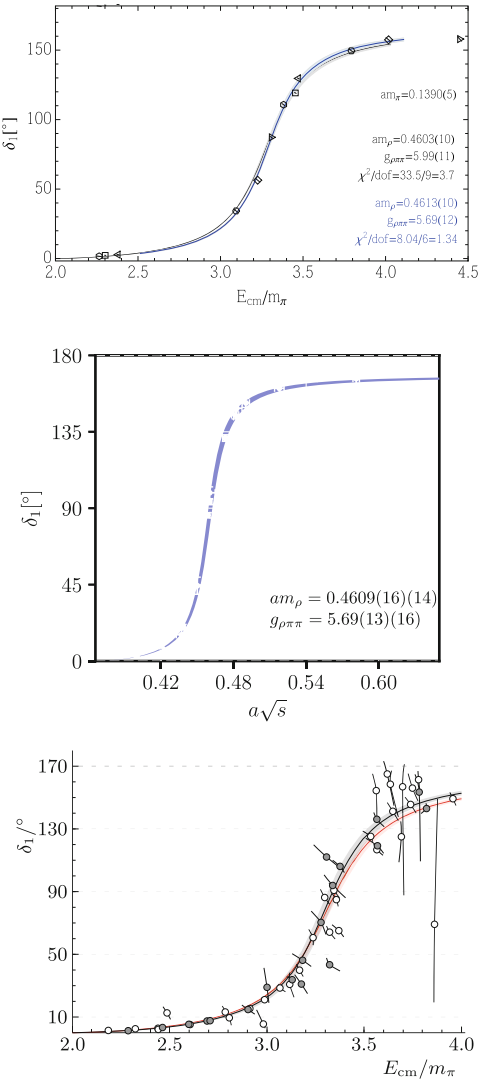


Fig. 2.5 Phase shifts for $\pi\pi$ scattering obtained for different pion masses. Top: reprinted with permission from [53] ($N_f = 2$, $M_\pi = 226$ MeV). © 2016, American Physical Society. All rights reserved. Middle: reprinted with permission from [54] ($N_f = 2 + 1$, $M_\pi = 316$ MeV). © 2017, American Physical Society. All rights reserved. Bottom: Reprinted under CC-BY-4.0 license from [55] ($N_f = 2 + 1$, $M_\pi = 220$ MeV). © 2019, The Author(s)

spread of data points. The error bars come from statistical uncertainties (not the L extrapolation). This is likely a good approximation, though the results are obtained for pions with roughly three times their physical mass, $M_\pi = 391$ MeV. Many collaborations repeated the same analysis for different values of the pion mass, see Fig. 2.5. These results extrapolate nicely to 5–10% for the ratio of the rho to pion mass.

Similar techniques can be applied to kaon-pion scattering, but this time with a new resonance called K^* [56]. This leads us to the most beautiful way to think about these resonances. The idea is to drive their mass so high that they become stable

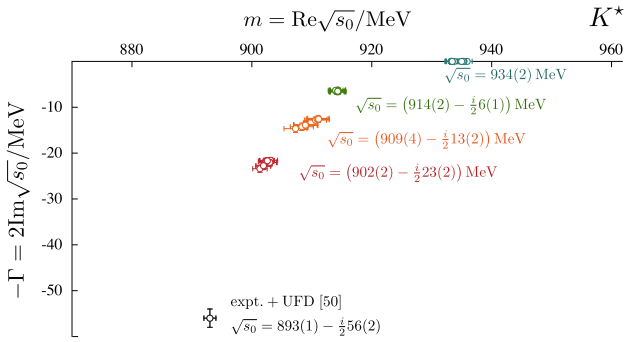


Fig. 2.6 Position of the complex pole associated with the K^* resonance in the complex energy $\sqrt{s_0}$-plane, for different values of the pion mass: $M_\pi = 391$ MeV (teal), $M_\pi = 327$ MeV (green), $M_\pi = 284$ MeV (orange), $M_\pi = 239$ MeV (red), and extrapolation to $M_\pi = 140$ MeV (black). Reprinted under CC-BY-4.0 license from [56]. © 2019, The Author(s)

states. You can then see how the pole in the complex plane moves from the real energy axis to right below the cut where it reaches its physical value. An example of this flow for the K^* resonance is shown in Fig. 2.6.

Audience Question 2.2.2. Can you reconstruct the tetraquark in the same way?
Answer: Yes, but the details may be different. An interesting example is the T_{cc} tetraquark candidate, which can be investigated in $DD^\star$ scattering, where the $D^\star$ is a stable particle in calculations where the pion mass is slightly heavier than at the physical point. For this channel a left-hand branchcut arises and when energies are extracted on the branchcut, an alternative formalism is needed. This is a very active area of research, see e.g. refs. [57–59].

Pretty much everything you can imagine to play with, people are already studying. For example, we can consider the D-meson scattering with a pion, which is a more complicated variation on the same underlying story. See [60] for the details of this case.

We now turn briefly to the complication of coupled channels. As mentioned before, putting the system in a box mixed different angular momenta. In practice, it means that the matrix F is fully populated. For example, for $K\pi \to K\pi$ scattering with $\mathbf{P} \neq 0$, we have to solve the system

$$\det\left[\begin{pmatrix} \mathcal{K}_s^{-1} & 0 \\ 0 & \mathcal{K}_p^{-1} \end{pmatrix} + \begin{pmatrix} F_{ss} & F_{sp} \\ F_{ps} & F_{pp} \end{pmatrix}\right] = 0\,, \tag{2.2.40}$$

where we have truncated the system to include only the s- and p-waves for the sake of illustration.

Similar issue happens when we have different flavor channels, except it is the $\mathcal{K}$-matrix that has the mixing terms. For example, if we denote $a = \pi\pi$ and $b = K\bar{K}$, then the equation reads

$$\det\left[\begin{pmatrix}\mathcal{K}_{a\to a} & \mathcal{K}_{a\to b}\\ \mathcal{K}_{b\to a} & \mathcal{K}_{b\to b}\end{pmatrix}^{-1} + \begin{pmatrix}F_a & 0\\ 0 & F_b\end{pmatrix}\right] = 0\,. \tag{2.2.41}$$

The goal is to make sure that the final result is as unbiased as possible by such issues. The workflow includes the following steps. First, consider correlators with a large operator basis, $\langle \mathcal{O}_a(\tau)\mathcal{O}_b^\dagger(0)\rangle$. Then, reliably extract finite-volume energies according to $\langle \Omega_m(\tau)\Omega_m^\dagger(0)\rangle \sim \mathrm{e}^{-E_m(L)\tau}$. By varying L and $\mathbf{P}$, we can recover a dense set of energies. On the other hand, we can identify a broad list of $\mathcal{K}$-matrix parametrizations: polynomials with poles, EFT-based ansatze, or dispersion-based approaches. All this can be used to make global fits to the finite-volume spectrum.

The literature of this field is vast and growing. Even restricting to numerical calculations in the past five years we find a huge number of papers [60–102]. The field is very active and there are many exciting results to come.

2.3 More Complicated Amplitudes

Maxwell T. Hansen

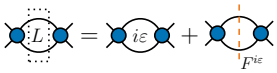
In the previous section, we have learned that observables computed on the lattice have a characteristic exponentially suppressed term. We used this suppression along with the finite lattice volume as a computational tool to extract $2 \to 2$ scattering amplitudes. In practice, we used a parametrization for the infinite-volume amplitude $\mathcal{M}$ in terms of the $\mathcal{K}$-matrix and fit the lattice data to it. The goal of this section is to move on to more aggressive applications in multi-particle observables.

We start by deriving a slightly modified version of the finite-volume correlator, using an $i\varepsilon$ prescription instead of a principal-valued one. We keep working in the regime of two-particle energies between $(2m)^2 < s < (4m)^2$. The starting point is the same formula as before for the finite-volume amplitude:

$$\mathcal{M}_L(P) = \quad + \quad + \quad + \cdots \tag{2.3.1}$$

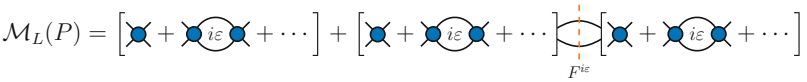
Recall that each dotted box contains a two-particle state of propagating pions in a finite-volume setup, and the blue blobs are the Bethe-Salpeter kernels from (2.2.18), which contain all contributions that are free of two-particle cuts. To get back the

infinite-volume amplitude, we break up each finite-volume loop into the infinite-volume one, plus a remainder $F^{i\varepsilon}$, defined by the equation



(2.3.2)

We can compare this with (2.2.36), where we broke the finite-volume loop into one with a principal-valued prescription instead of the $i\varepsilon$, and a remainder. Analogously to before, the dashed cut projects the loop to on-shell energies, and $F^{i\varepsilon}$ is a matrix of known geometric functions. Expanding out the series from (2.3.1) using (2.3.2) defines the finite-volume scattering amplitude in terms of the infinite-volume one,



$$= \frac{1}{\mathcal{M}(s)^{-1} + F^{i\varepsilon}(P,L)} \,. \tag{2.3.3}$$

Recall that the poles of $\mathcal{M}_L(P)$ give the finite-volume spectrum, and that the $i\varepsilon$ loop can be related to the sum over La. Since a sum cannot have branch cuts, the expression $\mathcal{M}_L$ has multiple poles that accumulate into a branch cut as $L \to \infty$.

Comparing with the previous subsection, we have now written the finite volume amplitude directly in terms of the infinite-volume one, $\mathcal{M}(s)$, instead of using the $\mathcal{K}$-matrix in intermediate steps. Although the two representations differ by a ρ term, they are completely equivalent: When we sum over the geometric series we get a result for $\mathcal{M}_L(P)$ that's independent of which representation we used in intermediate steps. The formulation with $\mathcal{M}(s)$ from (2.3.3) will be useful in what follows and provides a first step for using finite volume as a tool for more complicated amplitudes.

2.3.1 $1 + \mathcal{J} \to 2$

We first look at the process $1 + \mathcal{J} \to 2$, where $\mathcal{J}$ is a current. An example of such a process is the photo-production of a two-pion state from a single pion ($\pi\gamma \to \pi\pi$) or the analogous process for a kaon ($K\gamma \to K\pi$). The approach will be to use an expression analagous to (2.3.3) to relate such amplitudes to finite-volume matrix elements.

As a warm-up, note that one can rewrite (2.3.3) as

$$\mathcal{M}_L(P) = \mathcal{M}(s) - \mathcal{M}(s) \frac{1}{\mathcal{M}(s) + \left[F^{i\varepsilon}(P,L)\right]^{-1}} \mathcal{M}(s) \,. \tag{2.3.4}$$

This serves as inspiration for a similar decomposition of an alternative finite-volume correlator, denoted by $C_L(P)$ and defined schematically as

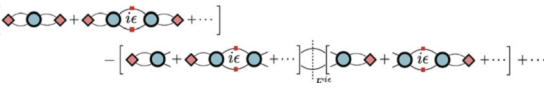
$$C_L(P) = \langle H|\mathcal{J}\mathcal{J}|H\rangle_L \,. \tag{2.3.5}$$

Here H stands for hadron and refers to a generic single-particle state. $\mathcal{J}$ indicates a generic local current, for example the electromagnetic current in the case of photo-production.

The idea is that the ket $\mathcal{J}|H\rangle_L$ and the bra ${}_L\langle H|\mathcal{J}$ correspond to creation and annihilation operators, respectively, that differ from those used to define $\mathcal{M}_L(P)$. However, the operators have the same internal quantum numbers, and thus this difference only modifies the leftmost and rightmost parts of the skeleton expansion, leaving the geometric series in the middle unchanged. Looking to (2.3.4) this means that the first term on the right-hand side will change, as will the outermost factors in the second term.

The full expression for C_L was derived in [41, 103, 104] and related work was also presented in [40,41,46,47,104–111]. The result can be summarized as follows:

$$C_L(P) = \langle H|\mathcal{J}\mathcal{J}|H\rangle_L \tag{2.3.6}$$



(2.3.7)

$$= C^{i\varepsilon}_{\infty}(s) - A^{i\varepsilon}_{\text{in}}(s)\frac{1}{\mathcal{M}(s) + [F^{i\varepsilon}(P,L)]^{-1}}A^{i\varepsilon}_{\text{out}}(s)\,. \tag{2.3.8}$$

In these diagrams, the blue circles are again the Bethe–Salpeter kernels and the red diamonds are the vertex functions from the new creation and annihilation operators. In the diagrammatic line we have split sums over two-particle loops into integrals, labeled $i\epsilon$ and sum-integral differences, labeled $F^{i\epsilon}$, as before. We have also grouped terms by the number of $F^{i\epsilon}$ factors. This is a useful representation for C_L because $A^{i\varepsilon}_{\text{in}}$ and $A^{i\varepsilon}_{\text{out}}$ are directly the physical infinite-volume amplitudes: $A^{i\varepsilon}_{\text{in},,\alpha} = \langle H|\mathcal{J}|\alpha, \text{in}\rangle$ and $A^{i\varepsilon}_{\text{out},,\alpha} = \langle \alpha, \text{out}|\mathcal{J}|H\rangle$, where α denotes all other quantum numbers of the two-particle state including angular momentum and flavour. (This follows diagramatically. Looking at the factor appearing to the) In (2.3.6) this index is contracted with the matrix appearing between the amplitudes, which is the same as the one appearing in (2.3.4).

Equation (2.3.6) has the same poles as Eq. (2.3.4) and therefore reveals nothing new about the relation between the finite-volume energies and the infinite-volume scattering amplitude. We do, however, uncover a new result by studying the residues at the poles. Focusing first on the left-hand side we note that

$$\lim_{E\to E_n(L)} [E - E_n(L)]\, C_L(P) = N(L)|\langle n|\mathcal{J}|H\rangle_L|^2 \,, \tag{2.3.9}$$

where $N(L)$ is a known normalization factor whose value depends on the details of the conventions used, see [103, 104]. The second quantity on the right-hand side is the squared magnitude of $\langle n|\mathcal{J}|H\rangle_L$, a finite-volume matrix element that can be computed on the lattice, where ${}_L\langle n|$ is a finite-volume excited state with the same quantum numbers as $\mathcal{J}|H\rangle_L$.

This same limit can be applied on the right-hand side of Eq. (2.3.6). Here we note that the pole does not appear in $C_\infty^{i\varepsilon}$, so the residue is given by the residue of the second term. To express this it is useful to define

$$\mathcal{R}(P,L) = -\lim_{E\to E_n(L)} \frac{E - E_n(L)}{\mathcal{M}(s) + \left[F^{i\varepsilon}(P,L)\right]^{-1}}\,. \tag{2.3.10}$$

Then, combining with Eq. (2.3.6), we find

$$N(L)|\langle n|\mathcal{J}|H\rangle_L|^2 = A_{\text{in}}^{i\varepsilon}(s)\mathcal{R}(P,L)A_{\text{out}}^{i\varepsilon}(s)\Big|_{s=P^2=(E_n(L),\boldsymbol{P})^2}\,. \tag{2.3.11}$$

The key idea is that the finite-volume matrix element on the left-hand side can be determined from numerical lattice QCD, and used to constrain the infinite-volume amplitudes $A_{\text{in}}^{i\varepsilon}(s)$ and $A_{\text{out}}^{i\varepsilon}(s)$ on the right-hand side.

One can additionally show that, once one projects to a given row of a given finite-volume irreducible representation, the matrix $\mathcal{R}$ is rank one. In particular, we can identify $\mu(E)$ as the eigenvalue of $\mathcal{M}(s) + [F^{i\varepsilon}(P,L)]^{-1}$ that vanishes at $E_n(L)$ and $\mathbf{v}$ as the corresponding normalized eigenvector. This allows us to write the residue as

$$\mathcal{R}(P,L) = \frac{1}{\mu'(E)}\mathbf{v}^T\mathbf{v}\,, \tag{2.3.12}$$

an effective relation on states can be written as

$$N(L)|\langle n|\mathcal{J}|H\rangle_L|^2 = \frac{1}{\mu'(E)}\left(\sum_{\alpha'}\langle H|\mathcal{J}|\alpha',\text{in}\rangle\mathbf{v}_{\alpha'}^T\right)\left(\sum_{\alpha}\mathbf{v}_\alpha\langle\alpha,\text{out}|\mathcal{J}|H\rangle\right)\,. \tag{2.3.13}$$

In addition, in certain cases it is possible to use the scattering matrix to rewrite the matrix element with the instate in terms of the matrix element with the outstate. Then, up to a phase ambiguity that must be treated with care [103, 104], (2.3.13) can be written as

$$\langle n|\mathcal{J}|H\rangle_L \propto \sum_\alpha \mathbf{v}_\alpha\langle\alpha,\text{out}|\mathcal{J}|H\rangle\,. \tag{2.3.14}$$

This gives a physical interpretation of the finite-volume matrix element in terms of linear combination of infinite-volume amplitudes, with relevant weights set by

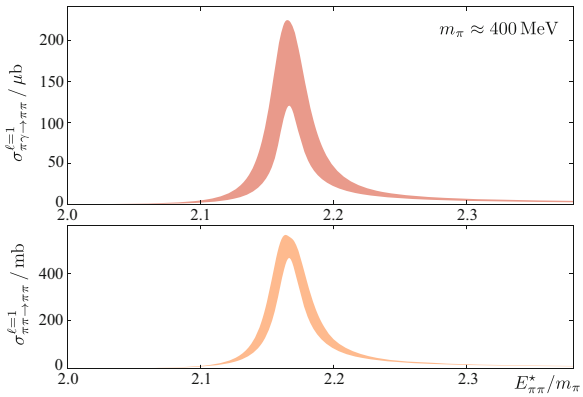
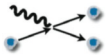


Fig. 2.7 The cross-section, proportional to $|\langle \pi\pi, \text{out}|\mathcal{J}_\mu|\pi\rangle|^2$ as a function of energy over the pion mass. The mass of the pion is given by $m_\pi \approx 400$ MeV. Reprinted with permission from [112]. © 2016, American Physical Society. All rights reserved

the eigenvector of the rank-one matrix $\mathcal{R}$. It is important to stress that not only the proportionality but indeed the full relation is known.

This result has been applied in numerical lattice QCD calculations in various cases in which only a single channel is present [40, 41, 46, 47, 104–111]. The first calculation involving an energy-infjecting current was for the amplitude $\pi\gamma \to \pi\pi$, with the final state in the channel of the ρ resonance, performed by the HadSpec collaboration in [112]. In this case the sum over infinite-volume matrix elements reduces to a single entry

$$\langle \pi\pi, \text{out}|\mathcal{J}_\mu|\pi\rangle \equiv$$  . (2.3.15)

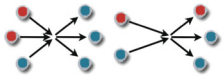
The results for the $\pi\gamma$ cross-section extracted in this work are shown in Fig. 2.7.

2.3.2 General Remarks on $2 + \mathcal{J} \to 2$, $2 \to 3$, and $3 \to 3$

We now turn to more complicated amplitudes involving three-particle states, either in the form of three asymptotic particles in an in or out state, or in the form of an external current that injects energy into a two-to-two system and mimics some of the features of a three-body amplitude.

There are multiple applications of three-hadron formalism in QCD, such as a systematic determination of pole positions, quantum numbers, and couplings, for resonances with a significant branching fraction to three particles. A few concrete examples include $\omega(782) \to \pi\pi\pi$, $a_1(1420) \to \pi\pi\pi$ and, $X(3872) \to J/\psi\pi\pi$. A long term goal is to reliably extract the properties of such resonances via lattice QCD, for various quark masses. Related applications include calculations of electroweak decay and transition amplitudes, three-body nuclear forces and structure observables, such as gluon content.

Focusing first on $2 \to 3$ and $3 \to 3$, the general goal is to have a finite-volume formalism, that also takes all constraints from S-matrix unitarity into account, for generic two- and three-particle systems, schematically represented by



(2.3.16)

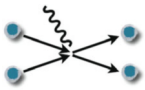
where the colors represent different species of particles. One generally aims to include any number of channels, identical and non-identical, and potentially non-degenerate particles, as well as particles with intrinsic spin. In fact by now these extensions have largely been accomplished [113–147].

The extension of the $2 \to 2$ formalism is not straightforward, mainly due to two key complications:

The first is that in higher-point processes, we have many more degrees of freedom than for $2 \to 2$. Concretely, a $2 \to 2$ scattering amplitude for spin-zero particles has 12 momentum components arising from four three-momenta where we intially ignore momentum conservation as well as rotational and Lorentz invariance. Ignoring these symmetries over counts by the 10 Poincaré generators (the four translations enforce energy and momentum conservation and additional redundances are removed by the three rotations and the three boosts). So, subtracting these we are left with $12 - 10 = 2$ degrees of freedom, which are the Mandelstam invariants s and t, or equivalently the CMF energy and the angle between incoming and outgoing particles in the CMF. Applying this logic to a $3 \to 3$ process of spin-zero particles, we start with 18 momentum components, and after subtracting the 10 Poincaré generators we still have 8 degrees of freedom. Similarly, a $2 \to 3$ of spin-zero particles process has 5 degrees of freedom.

The second complication comes from new possibilities of having classical on-shell propagating states. These configurations can cause singularities of the quantum field theory amplitude at the points where classical processes are kinematically allowed. Such singularities at classical kinematic configurations are called *Landau singularities*.

As a first step, we now consider the $2 + \mathcal{J} \to 2$ process. This is related to $2 \to 3$ but is defined with an external current (e.g. representing a photon) added to the initial or final state of a two-to-two amplitude. The target observable is the matrix element

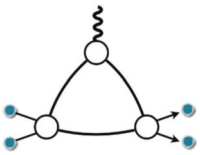
$$\langle \pi\pi, \text{out}|\mathcal{J}_\mu|\pi\pi, \text{in}\rangle \equiv$$ 

(2.3.17)

where the cartoon indicates the external legs.

As with the $2 \to 2$ and $1 + \mathcal{J} \to 2$ processes described above, the basic aprroach is to identify a finite-volume correlation function whose residues depend on finite-volume matrix elements related to $\langle \pi\pi, \text{out}|\mathcal{J}_\mu|\pi\pi, \text{in}\rangle$. The details of this are presented in [148–152] and for the present work we will only focus on one aspect

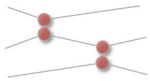
of this. Both the finite-volume correlator representing the $2 + \mathcal{J} \to 2$ process and the infinite-volume matrix element $\langle \pi\pi, \text{out}|\mathcal{J}_\mu|\pi\pi, \text{in}\rangle$ depend triangle diagrams of the form



(2.3.18)

That is, in the process $2 + \mathcal{J} \to 2$, we can have pairwise scattering for classical momenta of all the particles in the triangle diagram, including the internal ones. This causes an enhancement of the quantum amplitude at the corresponding phase space point; see Fig. 2.8 for a realization of a triangle singularity in a $2 + \mathcal{J} \to 2$ process. In general, it is difficult to disentangle Landau singularities, which cause an enhancement of the scattering amplitudes due to classical configurations, from enhancements coming from resonances.

In general, $3 \to 3$ scattering processes can also have Landau singularities due to the possibility of classical pairwise scatterings. In fact, we can solve for the classical-scattering configurations that are kinematically allowed, and thus count how many time such pairwise scatterings can occur. For example, if all particles in the scattering process have an equal mass, $m_1 = m_2 = m_3$, then we can work out that only up to three classical binary collisions are possible. Diagrammatically, this leads to two types of allowed pairwise scatterings (up to permutations), the first being a pole in a configuration of two binary collisions,



(2.3.19)

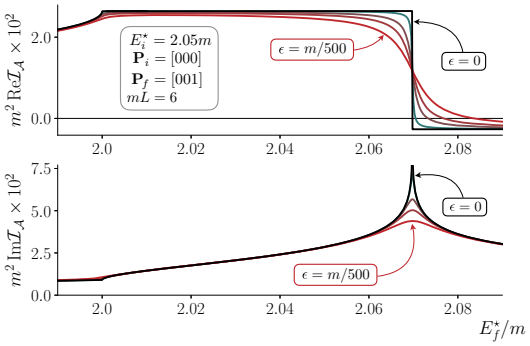
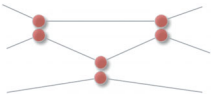


Fig. 2.8 The amplitude for $2 + \mathcal{J} \to 2$, here $\pi\pi \to J\pi\pi$, can be split into a part that can be evaluated analytically, $\mathcal{I}_A$, and a remainder that can be evaluated numerically. One can then deduce the Landau singularities of $\mathcal{I}_A$ by computing the relevant Feynman integrals. **Top:** The real part of $\mathcal{I}_A$, which is discontinuous due to the triangle Landau singularity. **Bottom:** The imaginary part of $\mathcal{I}_A$, which is singular due to the triangle Landau singularity. Reprinted under CC-BY-4.0 license from [149]. © 2019, The Author(s)

while the second is a triangle-diagram configuration:



(2.3.20)

If the masses are not equal, we open up further possibilities for allowed pairwise scatterings. If at least one of the masses is lower than the other two, e.g. if the masses satisfy $m_1 = m_2 = m_3 - \epsilon$, then 4 collisions are possible. This would be the case, for example, in $\pi\pi K \to \pi\pi K$ scattering. The general result from [153] is that if we define the mass ratio

$$b = \frac{(m_1 + m_3)(m_2 + m_3)}{m_1 m_2}\,, \tag{2.3.21}$$

then $2n + 1$ collisions or more are disallowed, where n satisfies

$$b^{n-2}(b-1) > 1\,. \tag{2.3.22}$$

This shows that at most 3 binary collisions are possible for equal masses, and when $b < 2$, which is e.g. the case in $\pi K K$ scattering, only up to 5 binary collisions are possible.

Audience Question 2.3.1. Was this relation not only for non-relativistic scattering? *Answer:* Yes, but we are in the non-relativistic regime here.

There are two key implementations we use to incorporate Landau singularities into the lattice formalism. The first one is to define a modified $\mathcal{K}$-matrix obtained by subtracting the singular contributions,

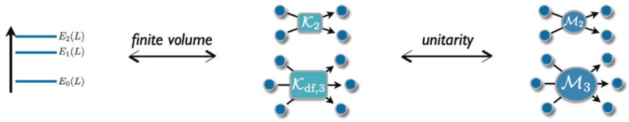
$$\mathcal{K}_{\mathrm{df},3} = \begin{matrix}\text{fully connected diagrams}\\ \text{with PV pole prescription}\end{matrix} - \left(\quad + \quad + \cdots \right)\,. \tag{2.3.23}$$

We have added the subscript "df" to emphasize that this new $\mathcal{K}$-matrix is "divergence free". In fact, it's a smooth, real-valued function. Note that we generally include all connected diagrams in the brackets, anticipating arbitrary masses. The second one is to use that $\mathcal{K}_{\mathrm{df},3}$ has a systematic low-energy expansion,

$$\mathcal{K}_{\mathrm{df},3}(p_3, p_2, p_1; k_3, k_2, k_1) = \mathcal{K}_{\mathrm{df},3}^{\mathrm{iso},0} + \mathcal{K}_{\mathrm{df},3}^{\mathrm{iso},1}\Delta + \ldots\,, \tag{2.3.24}$$

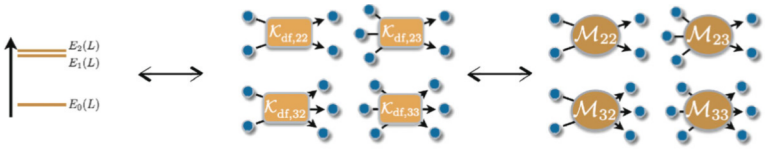
where $\Delta = \frac{s-(3m)^2}{(3m)^2}$. This expansion is analogous to an effective range expansion and gives a handle on many degrees of freedom, which enter order by order.

We can now derive general relations between energies and two-and-three scalar scattering [114–116]. If there is no $2 \to 3$ scattering, we can directly use the divergence-free kernel for $2 \to 2$ and $3 \to 3$ processes,



(2.3.25)

However, if $2 \to 3$ is an allowed process in our theory, we are forced to include the $2 \to 3$ divergence-free kernel, even if we are only interested in amplitudes of the form $2 \to 2$ or $3 \to 3$. This is because the sub-channel always appears between intermediate states,



(2.3.26)

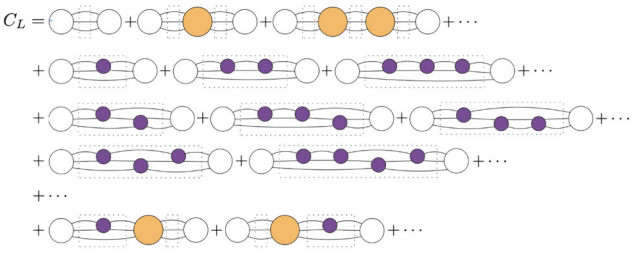
Including sub-channel resonances, different isospins and assuming non-degenerate configurations, we can, for example, get a $\pi\pi\pi \to \pi\pi\pi$ process of the type

$$\pi\pi\pi \to \rho\pi \to \omega \to \rho\pi \to \pi\pi\pi \,. \tag{2.3.27}$$

Note that the discussion in this subsection focused on the region $(2m)^2 < s < (4m)^2$, and, so far, computations have only been performed for those kinematics.

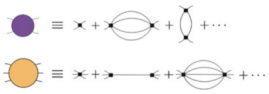
2.3.3 3-Particle Derivation

We now turn to $3 \to 3$ scattering. It turns out that the methods for treating $2 \to 2$ scattering generalize to $3 \to 3$ processes. We start with studying the 3-body correlator in an all-orders skeleton expansion [114],



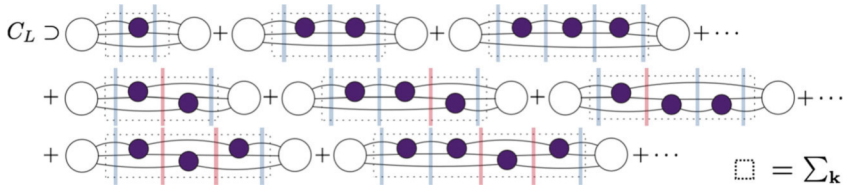
(2.3.28)

where the circles denote kernels for $2 \to 2$ and $3 \to 3$ processes; diagrammatically,



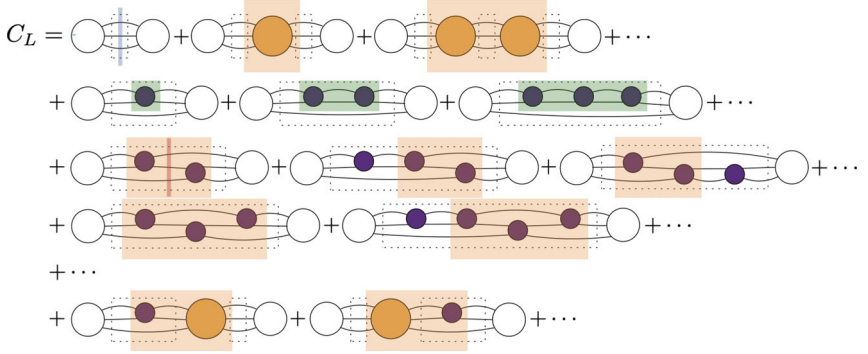
(2.3.29)

We have dropped the L dependence of the kernels which scales as e^{-mL}, and the solid lines denote fully dressed hadrons as before. We distinguish two types of cuts, as represented diagrammatically by the following red and blue shadings:



(2.3.30)

The blue cuts will be denoted below by $\mathbf{F}$ and the red ones are denoted with $\mathbf{G}$. Note that this is not quite the final answer for C_L; we're missing the short-distance contributions from off-shellness and smooth terms represented by the divergence-free kernels. To fix this, we include the short-distance-parts and summation as



(2.3.31)

Here, the orange shading denotes contributions to $\mathbf{K}_{\mathrm{df},3}$ and the green denotes contributions to $\mathbf{K}_2$. In analogy to the $2 \to 2$ case, we can now write a relation between finite-volume correlator and the infinite-volume one,

$$C_L - C_\infty = \mathbf{A}_3' \mathbf{F}_{33} \mathbf{A}_3 + \mathbf{A}_3' \mathbf{F}_{33} \mathbf{K}_{\mathrm{df},3} \mathbf{F}_{33} \mathbf{A}_3 + \ldots = \mathbf{A}_3' \frac{1}{\mathbf{F}_{33}^{-1} + \mathbf{K}_{\mathrm{df},3}} \mathbf{A}_3 \,, \tag{2.3.32}$$

with

$$\mathbf{F}_{33} = \frac{1}{3}\mathbf{F} + \mathbf{F}\mathbf{K}_2 \frac{1}{1 - (\mathbf{F} + \mathbf{G})\mathbf{K}_2}\mathbf{F}\,. \tag{2.3.33}$$

This last equation incorporates all the different contributions, from both long and short distance.

The poles in C_L correspond to the finite-volume energies. These occur whenever

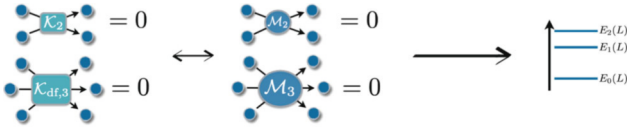
$$\det\left[\mathbf{K}_{\mathrm{df},3}^{-1} + \mathbf{F}_{33}\right] = 0. \tag{2.3.34}$$

This relation holds for three-particle energies, but neglects terms of order e^{-mL}. See Refs. [114, 115, 121, 154–157] for more details, and [158] for a review.

2.3.4 Examples

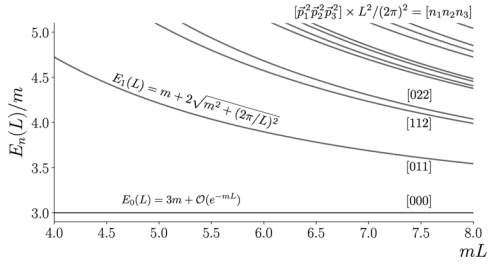
Now that we have developed the relevant tools, we can start applying this formalism to different examples.

As a warm-up example, let's derive an energy spectrum for non-interacting energies, i.e.



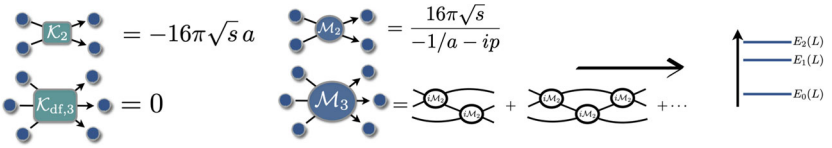
(2.3.35)

which gives an energy spectrum as a function of mL,



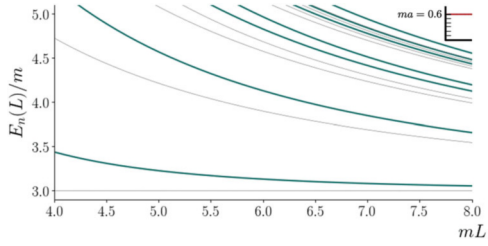
(2.3.36)

Next, let's add interactions only, but assume that the divergence-free $3 \to 3$ kernel vanishes,



(2.3.37)

This construction leads to a different energy spectrum, for example for $ma = 0.6$ we have



(2.3.38)

In general, these toy examples are useful for isolating how different terms and contributions affect the energy spectrum [122].

Next, let us move onto physical processes such as $\pi^+\pi^+\pi^+ \to \pi^+\pi^+\pi^+$. The workflow starts with determining the finite-volume energy spectrum. Next, we use a fit to constrain the two- and three-body $\mathcal{K}$-matrices, since they can be related to the energies through the finite-volume formalism. The technical strategy is to fit both two and three-body $\mathcal{K}$ to various polynomials, and put a cut on the CMF energy in the fits. Note that $\mathcal{K}_{\text{df},3}$ is scheme dependent, but this dependence is drops out in the amplitude. The final step is to use unitarity to solve integral equations to extract $3 \to 3$ scattering. Some of the different ingredients entering the computation are shown in Fig. 2.9.

These results can now be combined into computing the three-particle amplitude, with some results shown in Fig. 2.10.

2.4 Conclusions

Lattice QCD is in the era of rigorous resonance spectroscopy. In these lectures, we have seen how finite volume is a useful tool for precision extractions of QCD observables, with examples ranging from two-point functions to $3 \to 3$ scattering processes.

There are many directions on which further progress is currently being made. First, although the ground work has been set on the formal analysis side, further analysis of physical scattering processes is still ongoing. Second, progress has been

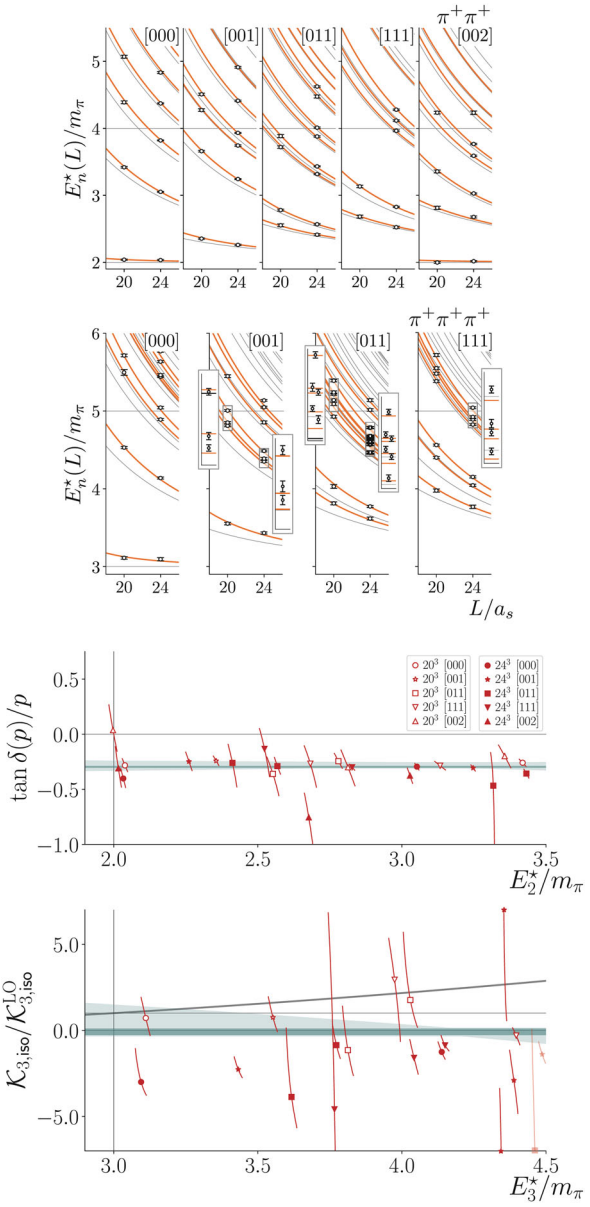


Fig. 2.9 The energy spectrum for $\pi^+\pi^+$ (**1st panel**) and $\pi^+\pi^+\pi^+$ (**2nd panel**), and examples of fits to the $\mathcal{K}$ matrix. The two final panels show the lattice QCD data as black points. The orange curves show a single parameter fit to the data. The right panel shows the K-matrices resulting from these fits. The points in the two final panels show the result of directly mapping a given finite-volume energy to a value of K at a given energy. Reprinted under CC-BY-4.0 license from [64]. © 2021, The Author(s)

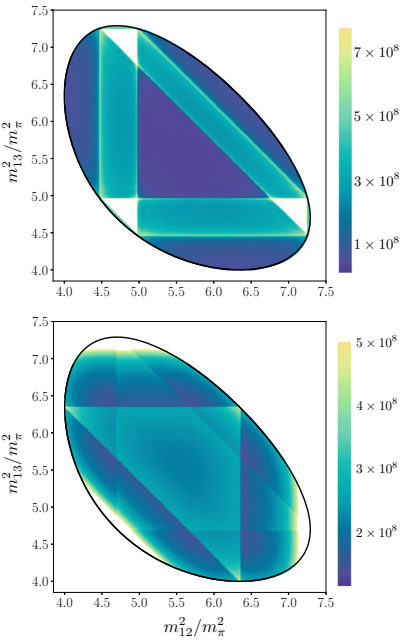


Fig. 2.10 The three-particle amplitude $m_\pi^4 |\mathcal{M}_3|$ for $\pi^+\pi^+\pi^+$ scattering. The CMF energy is $\sqrt{s_3} = 3.7m$, and the kinematics are fixed to certain values. Reprinted under CC-BY-4.0 license from [64]. © 2021, The Author(s)

made recently on advanced algorithms that enables high precision for the excited states, which is important to incorporate for precision scattering on a lattice. Third, using physical masses on the lattice are now becoming increasingly tractable, while previous calculations have often used unphysical quark masses. More generally, varying the masses probes the resonance structure of the theory.

Some of the next steps forward in this field are to complete the 3-particle formalism and extend it to an N-particle one. We would also like to extend the studies involving an external current to higher points, and push more channels into the precision regime. In summary, the field is active with many more results expected in the near future.

References

1. Flavour Lattice Averaging Group (FLAG) Collaboration, Y. Aoki, et al., *FLAG Review 2024* 2411.04268
2. ETM Collaboration, N. Carrasco et al., B-Physics from $N_f = 2$ tmQCD: the standard model and beyond. J. High Energy Phys. **03**, 016 (2014). [1308.1851]
3. ALPHA Collaboration, F. Bernardoni et al., Decay constants of B-mesons from non-perturbative HQET with two light dynamical quarks. Phys. Lett. B **735**, 349 (2014). [1404.3590]
4. Fermilab Lattice, MILC Collaboration, A. Bazavov et al., B- and D-meson decay constants from three-flavor lattice QCD. Phys. Rev. D **85**, 114506 (2012). [1112.3051]
5. C. McNeile, C.T.H. Davies, E. Follana, K. Hornbostel, G.P. Lepage, High-precision f_{B_s} and HQET from relativistic lattice QCD. Phys. Rev. D **85**, 031503 (2012). [1110.4510]

6. H. Na, C.J. Monahan, C.T.H. Davies, R. Horgan, G.P. Lepage, J. Shigemitsu, The B and B_s meson decay constants from lattice QCD. Phys. Rev. D **86**, 034506 (2012). [1202.4914]
7. Y. Aoki, T. Ishikawa, T. Izubuchi, C. Lehner, A. Soni, Neutral B meson mixings and B meson decay constants with static heavy and domain-wall light quarks. Phys. Rev. D **91**, 114505 (2015). [1406.6192]
8. N.H. Christ, J.M. Flynn, T. Izubuchi, T. Kawanai, C. Lehner, A. Soni, R.S. Van de Water, O. Witzel, B-meson decay constants from 2+1-flavor lattice QCD with domain-wall light quarks and relativistic heavy quarks. Phys. Rev. D **91**, 054502 (2015). [1404.4670]
9. HPQCD Collaboration, R.J. Dowdall, C.T.H. Davies, R.R. Horgan, C.J. Monahan, J. Shigemitsu, B-meson decay constants from improved lattice nonrelativistic QCD with physical u, d, s, and c quarks. Phys. Rev. Lett. **110**, 222003 (2013). [1302.2644]
10. ETM Collaboration, A. Bussone et al., Mass of the b quark and B -meson decay constants from N_f=2+1+1 twisted-mass lattice QCD. Phys. Rev. D **93**, 114505 (2016). [1603.04306]
11. C. Hughes, C.T.H. Davies, C.J. Monahan, New methods for B meson decay constants and form factors from lattice NRQCD. Phys. Rev. D **97**, 054509 (2018). [1711.09981]
12. A. Bazavov et al., B- and D-meson leptonic decay constants from four-flavor lattice QCD. Phys. Rev. D **98**, 074512 (2018). [1712.09262]
13. R. Balasubramamian, B. Blossier, Decay constant of B_s and B_s^* mesons from $N_f = 2$ lattice QCD. Eur. Phys. J. C **80**, 412 (2020). [1912.09937]
14. C.T.H. Davies, C. McNeile, E. Follana, G.P. Lepage, H. Na, J. Shigemitsu, Update: precision D_s decay constant from full lattice QCD using very fine lattices. Phys. Rev. D **82**, 114504 (2010). [1008.4018]
15. P.A. Boyle, L. Del Debbio, A. Jüttner, A. Khamseh, F. Sanfilippo, J.T. Tsang, The decay constants $\mathbf{f_D}$ and $\mathbf{f_{D_s}}$ in the continuum limit of $\mathbf{N_{f=2+1}}$ domain wall lattice QCD. J. High Energy Phys. **12**, 008 (2017). [1701.02644]
16. Y.B. Yang et al., Charm and strange quark masses and f_{D_s} from overlap fermions. Phys. Rev. D **92**, 034517 (2015). [1410.3343]
17. H. Na, C.T.H. Davies, E. Follana, G.P. Lepage, J. Shigemitsu, $|V_{cd}|$ from D meson leptonic decays. Phys. Rev. D **86**, 054510 (2012). [1206.4936]
18. N. Carrasco et al., Leptonic decay constants f_K, f_D, and f_{D_s} with $N_f = 2 + 1 + 1$ twisted-mass lattice QCD. Phys. Rev. D **91**, 054507 (2015). [1411.7908]
19. M. Luscher, Volume dependence of the energy spectrum in massive quantum field theories. 1. Stable particle states. Commun. Math. Phys. **104**, 177 (1986)
20. M.T. Hansen, A. Patella, Finite-volume effects in $(g-2)_\mu^{\mathrm{HVP,LO}}$. Phys. Rev. Lett. **123**, 172001 (2019). [1904.10010]
21. M.T. Hansen, A. Patella, Finite-volume and thermal effects in the leading-HVP contribution to muonic $(g-2)$. J. High Energy Phys. **10**, 029 (2020). [2004.03935]
22. T. Aoyama, et al., The anomalous magnetic moment of the muon in the Standard Model. Phys. Rep. **887**, 1 (2020). [2006.04822]
23. D. Bernecker, H.B. Meyer, Vector correlators in lattice QCD: methods and applications. Eur. Phys. J. A **47**, 148 (2011). [1107.4388]
24. R.F. Lebed, R.E. Mitchell, E.S. Swanson, Heavy-quark QCD exotica. Prog. Part. Nucl. Phys. **93**, 143 (2017). [1610.04528]
25. LHCb Collaboration, R. Aaij et al., Quantum numbers of the $X(3872)$ state and orbital angular momentum in its $\rho^0 J\psi$ decay. Phys. Rev. D **92**, 011102 (2015). [1504.06339]
26. LHCb Collaboration, R. Aaij et al., Observation of CP violation in charm decays. Phys. Rev. Lett. **122**, 211803 (2019). [1903.08726]
27. A. Soni, Resonance enhancement of Charm CP. 1905.00907
28. S.D. Protopopescu, M. Alston-Garnjost, A. Barbaro-Galtieri, S.M. Flatte, J.H. Friedman, T.A. Lasinski, G.R. Lynch, M.S. Rabin, F.T. Solmitz, Pi pi partial wave analysis from reactions pi+ p —> pi+ pi- Delta++ and pi+ p —> K+ K- Delta++ at 7.1-GeV/c. Phys. Rev. D **7**, 1279 (1973)
29. L. Maiani, M. Testa, Final state interactions from Euclidean correlation functions. Phys. Lett. B **245**, 585 (1990)

30. M.T. Hansen, H.B. Meyer, D. Robaina, From deep inelastic scattering to heavy-flavor semileptonic decays: total rates into multihadron final states from lattice QCD. Phys. Rev. D **96**, 094513 (2017). [1704.08993]
31. J. Bulava, M.T. Hansen, Scattering amplitudes from finite-volume spectral functions. Phys. Rev. D **100**, 034521 (2019). [1903.11735]
32. M. Hansen, A. Lupo, N. Tantalo, Extraction of spectral densities from lattice correlators. Phys. Rev. D **99**, 094508 (2019). [1903.06476]
33. M. Bruno, M.T. Hansen, Variations on the Maiani-Testa approach and the inverse problem. J. High Energy Phys. **06**, 043 (2021). [2012.11488]
34. A. Patella, N. Tantalo, Scattering amplitudes from Euclidean correlators: Haag-Ruelle theory and approximation formulae. J. High Energy Phys. **01**, 091 (2025). [2407.02069]
35. M. Luscher, Volume dependence of the energy spectrum in massive quantum field theories. 2. Scattering states. Commun. Math. Phys. **105**, 153 (1986)
36. M. Luscher, Two particle states on a torus and their relation to the scattering matrix. Nucl. Phys. B **354**, 531 (1991)
37. K. Rummukainen, S.A. Gottlieb, Resonance scattering phase shifts on a nonrest frame lattice. Nucl. Phys. B **450**, 397 (1995). [hep-lat/9503028]
38. X. Feng, X. Li, C. Liu, Two particle states in an asymmetric box and the elastic scattering phases. Phys. Rev. D **70**, 014505 (2004). [hep-lat/0404001]
39. S. He, X. Feng, C. Liu, Two particle states and the S-matrix elements in multi-channel scattering. J. High Energy Phys. **07**, 011 (2005). [hep-lat/0504019]
40. N.H. Christ, C. Kim, T. Yamazaki, Finite volume corrections to the two-particle decay of states with non-zero momentum. Phys. Rev. D **72**, 114506 (2005). [hep-lat/0507009]
41. C.H. Kim, C.T. Sachrajda, S.R. Sharpe, Finite-volume effects for two-hadron states in moving frames. Nucl. Phys. B **727**, 218 (2005). [hep-lat/0507006]
42. M. Lage, U.-G. Meissner, A. Rusetsky, A method to measure the antikaon-nucleon scattering length in lattice QCD. Phys. Lett. B **681**, 439 (2009). [0905.0069]
43. V. Bernard, M. Lage, U.G. Meissner, A. Rusetsky, Scalar mesons in a finite volume. J. High Energy Phys. **01**, 019 (2011). [1010.6018]
44. Z. Fu, Rummukainen-Gottlieb's formula on two-particle system with different mass. Phys. Rev. D **85**, 014506 (2012). [1110.0319]
45. L. Leskovec, S. Prelovsek, Scattering phase shifts for two particles of different mass and non-zero total momentum in lattice QCD. Phys. Rev. D **85**, 114507 (2012). [1202.2145]
46. R.A. Briceno, Z. Davoudi, Moving multichannel systems in a finite volume with application to proton-proton fusion. Phys. Rev. D **88**, 094507 (2013). [1204.1110]
47. M.T. Hansen, S.R. Sharpe, Multiple-channel generalization of Lellouch-Luscher formula. Phys. Rev. D **86**, 016007 (2012). [1204.0826]
48. P. Guo, J. Dudek, R. Edwards, A.P. Szczepaniak, Coupled-channel scattering on a torus. Phys. Rev. D **88**, 014501 (2013). [1211.0929]
49. R.A. Briceno, Z. Davoudi, T.C. Luu, Two-nucleon systems in a finite volume: (I) quantization conditions. Phys. Rev. D **88**, 034502 (2013). [1305.4903]
50. R.A. Briceno, Two-particle multichannel systems in a finite volume with arbitrary spin. Phys. Rev. D **89**, 074507 (2014). [1401.3312]
51. K. Huang, C.N. Yang, Quantum-mechanical many-body problem with hard-sphere interaction. Phys. Rev. **105**, 767 (1957)
52. Hadron Spectrum Collaboration, J.J. Dudek, R.G. Edwards, C.E. Thomas, Energy dependence of the ρ resonance in $\pi\pi$ elastic scattering from lattice QCD. Phys. Rev. D **87**, 034505 (2013). [1212.0830]
53. D. Guo, A. Alexandru, R. Molina, M. Döring, Rho resonance parameters from lattice QCD. Phys. Rev. D **94**, 034501 (2016). [1605.03993]
54. C. Alexandrou, L. Leskovec, S. Meinel, J. Negele, S. Paul, M. Petschlies, A. Pochinsky, G. Rendon, S. Syritsyn, P-wave $\pi\pi$ scattering and the ρ resonance from lattice QCD. Phys. Rev. D **96**, 034525 (2017). [1704.05439]

55. C. Andersen, J. Bulava, B. Hörz, C. Morningstar, The $I = 1$ pion-pion scattering amplitude and timelike pion form factor from $N_f = 2 + 1$ lattice QCD. Nucl. Phys. B **939**, 145 (2019). [1808.05007]
56. D.J. Wilson, R.A. Briceno, J.J. Dudek, R.G. Edwards, C.E. Thomas, The quark-mass dependence of elastic πK scattering from QCD. Phys. Rev. Lett. **123**, 042002 (2019). [1904.03188]
57. A.B.A. Raposo, M.T. Hansen, Finite-volume scattering on the left-hand cut. J. High Energy Phys. **08**, 075 (2024). [2311.18793]
58. M.T. Hansen, F. Romero-López, S.R. Sharpe, Incorporating DDπ effects and left-hand cuts in lattice QCD studies of the $T_{cc}(3875)^+$. J. High Energy Phys. **06**, 051 (2024). [2401.06609]
59. A.B.A. Raposo, R.A. Briceño, M.T. Hansen, A.W. Jackura, Extracting scattering amplitudes for arbitrary two-particle systems with one-particle left-hand cuts via lattice QCD (2025). 2502.19375
60. Hadron Spectrum Collaboration, L. Gayer, N. Lang, S. M. Ryan, D. Tims, C.E. Thomas, D.J. Wilson, Isospin-1/2 Dπ scattering and the lightest D_0^* resonance from lattice QCD. J. High Energy Phys. **07**, 123 (2021). [2102.04973]
61. Extended Twisted Mass, ETM Collaboration, M. Fischer, B. Kostrzewa, M. Mai, M. Petschlies, F. Pittler, M. Ueding, C. Urbach, M. Werner, The ρ-resonance from Nf=2 lattice QCD including the physical pion mass. Phys. Lett. B **819**, 136449 (2021). [2006.13805]
62. Hadron Spectrum Collaboration, G.K.C. Cheung, C.E. Thomas, D.J. Wilson, G. Moir, M. Peardon, S.M. Ryan, DK I = 0,$D\overline{K}$I = 0, 1 scattering and the $D_{s0}^*(2317)$ from lattice QCD. J. High Energy Phys. **02**, 100 (2021). [2008.06432]
63. Hadron Spectrum Collaboration, A.J. Woss, J.J. Dudek, R.G. Edwards, C.E. Thomas, D.J. Wilson, Decays of an exotic $1-+$ hybrid meson resonance in QCD. Phys. Rev. D **103**, 054502 (2021). [2009.10034]
64. Hadron Spectrum Collaboration, M.T. Hansen, R.A. Briceño, R.G. Edwards, C.E. Thomas, D.J. Wilson, Energy-dependent $\pi^+\pi^+\pi^+$ Scattering Amplitude from QCD. Phys. Rev. Lett. **126** 012001 (2021). [2009.04931]
65. Hadron Spectrum Collaboration, C.T. Johnson, J.J. Dudek, Excited J^{--} meson resonances at the SU(3) flavor point from lattice QCD. Phys. Rev. D **103**, 074502 (2021). [2012.00518]
66. J.R. Green, A.D. Hanlon, P.M. Junnarkar, H. Wittig, Weakly bound H dibaryon from SU(3)-flavor-symmetric QCD. Phys. Rev. Lett. **127**, 242003 (2021). [2103.01054]
67. T.D. Blanton, A.D. Hanlon, B. Hörz, C. Morningstar, F. Romero-López, S.R. Sharpe, Interactions of two and three mesons including higher partial waves from lattice QCD. J. High Energy Phys. **10**, 023 (2021). [2106.05590]
68. GWQCD Collaboration, M. Mai, A. Alexandru, R. Brett, C. Culver, M. Döring, F.X. Lee, D. Sadasivan, Three-body dynamics of the a1(1260) resonance from lattice QCD. Phys. Rev. Lett. **127**, 222001 (2021). [2107.03973]
69. C. Morningstar, J. Bulava, A.D. Hanlon, B. Hörz, D. Mohler, A. Nicholson, S. Skinner, A. Walker-Loud, Progress on Meson-Baryon scattering. PoS **LATTICE2021**, 170 (2022). [2111.07755]
70. P. Madanagopalan, J. Bulava, J.R. Green, A.D. Hanlon, B. Hörz, P. Junnarkar, C. Morningstar, S. Paul, H. Wittig, H dibaryon away from the $SU(3)_f$ symmetric point. PoS **LATTICE2021**, 459 (2022). [2111.11541]
71. J.R. Green, A.D. Hanlon, P.M. Junnarkar, H. Wittig, Continuum limit of baryon-baryon scattering with SU(3) flavor symmetry. PoS **LATTICE2021**, 294 (2022). [2111.09675]
72. M. Padmanath, S. Prelovsek, Signature of a doubly charm tetraquark pole in DD* scattering on the lattice. Phys. Rev. Lett. **129**, 032002 (2022). [2202.10110]
73. Hadron Spectrum Collaboration, A. Radhakrishnan, J.J. Dudek, R.G. Edwards, Radiative decay of the resonant K* and the γK $\rightarrow$Kπ amplitude from lattice QCD. Phys. Rev. D **106**, 114513 (2022). [2208.13755]
74. Hadron Spectrum Collaboration, N. Lang, D.J. Wilson, Axial-vector D1 hadrons in D*π scattering from QCD. Phys. Rev. Lett. **129**, 252001 (2022). [2205.05026]

75. J. Bulava, A.D. Hanlon, B. Hörz, C. Morningstar, A. Nicholson, F. Romero-López, S. Skinner, P. Vranas, A. Walker-Loud, Elastic nucleon-pion scattering at mπ = 200 MeV from lattice QCD. Nucl. Phys. B **987**, 116105 (2023). [2208.03867]
76. Baryon Scattering (BaSc) Collaboration, J.R. Green, A.D. Hanlon, P.M. Junnarkar, H. Wittig, Nucleon-nucleon scattering from distillation. PoS **LATTICE2022**, 200 (2023). [2212.09587]
77. J. Bulava, B. Hörz, Renwick, J. Hudspdith, C. Johnson, D. Mohler, C. Morningstar, D meson - pion scattering on CLS 2+1 flavor ensembles. PoS **LATTICE2022**, 068 (2023)
78. Y. Lyu, S. Aoki, T. Doi, T. Hatsuda, Y. Ikeda, J. Meng, Doubly charmed tetraquark Tcc+ from lattice QCD near physical point. Phys. Rev. Lett. **131**, 161901 (2023). [2302.04505]
79. Z.T. Draper, A.D. Hanlon, B. Hörz, C. Morningstar, F. Romero-López, S.R. Sharpe, Interactions of πK, $\pi\pi$K and KKπ systems at maximal isospin from lattice QCD. J. High Energy Phys. **05**, 137 (2023). [2302.13587]
80. Hadron Spectrum Collaboration, A. Rodas, J.J. Dudek, R.G. Edwards, Quark mass dependence of $\pi\pi$ scattering in isospin 0, 1, and 2 from lattice QCD. Phys. Rev. D **108**, 034513 (2023). [2303.10701]
81. Hadron Spectrum Collaboration, A. Rodas, J.J. Dudek, R.G. Edwards, Determination of crossing-symmetric $\pi\pi$ scattering amplitudes and the quark mass evolution of the σ constrained by lattice QCD. Phys. Rev. D **109**, 034513 (2024). [2304.03762]
82. B. Scattering (BaSc) Collaboration, J. Bulava et al., Lattice QCD study of $\pi\Sigma$-K$^-$N scattering and the $\Lambda(1405)$ resonance. Phys. Rev. D **109**, 014511 (2024). [2307.13471]
83. Baryon Scattering (BaSc) Collaboration, J. Bulava et al., Two-Pole nature of the $\Lambda(1405)$ resonance from lattice QCD. Phys. Rev. Lett. **132**, 051901 (2024). [2307.10413]
84. Hadron Spectrum Collaboration, D.J. Wilson, C.E. Thomas, J.J. Dudek, R.G. Edwards, Charmonium χc0 and χc2 resonances in coupled-channel scattering from lattice QCD. Phys. Rev. D **109**, 114503 (2024). [2309.14071]
85. Hadron Spectrum Collaboration, D.J. Wilson, C.E. Thomas, J.J. Dudek, R.G. Edwards, Scalar and tensor charmonium resonances in coupled-channel scattering from lattice QCD. Phys. Rev. Lett. **132**, 241901 (2024). [2309.14070]
86. J. Bulava, et al., Low-lying baryon resonances from lattice QCD. Nuovo Cim. C **47**, 165 (2024). [2310.08375]
87. S. Skinner, J. Bulava, D. Darvish, A.D. Hanlon, B. Hoerz, C. Morningstar, A. Nicholson, F. Romero-López, P. Vranas, A. Walker-Loud, Lattice QCD studies of the Δ baryon resonance and the $K_0^*(700)$ and $a_0(980)$ meson resonances: the role of exotic operators in determining the finite-volume spectrum. PoS **LATTICE2023**, 074 (2024). [2312.10184]
88. J. Bulava et al., The $\Lambda(1405)$ pole structure from Lattice QCD: a coupled-channel $\pi\Sigma-$ KN study. EPJ Web Conf. **303**, 01004 (2024)
89. S. Collins, A. Nefediev, M. Padmanath, S. Prelovsek, Toward the quark mass dependence of Tcc+ from lattice QCD. Phys. Rev. D **109**, 094509 (2024). [2402.14715]
90. Hadron Spectrum Collaboration, J.D.E. Yeo, C.E. Thomas, D.J. Wilson, DK/Dπ scattering and an exotic virtual bound state at the SU(3) flavour symmetric point from lattice QCD. J. High Energy Phys. **07**, 012 (2024). [2403.10498]
91. P. Boyle, F. Erben, V. Gülpers, M.T. Hansen, F. Joswig, N.P. Lachini, M. Marshall, A. Portelli, Light and strange vector resonances from lattice QCD at physical quark masses (2024). 2406.19194
92. Hadron Spectrum Collaboration, T. Whyte, D.J. Wilson, C.E. Thomas, Near-threshold states in coupled DD*-D*D* scattering from lattice QCD. Phys. Rev. D **111**, 034511 (2025). [2405.15741]
93. Hadron Spectrum Collaboration, J.J. Dudek, C.T. Johnson, Coupled-channel J– meson resonances from lattice QCD. Phys. Rev. D **110**, 034512 (2024). [2406.07261]
94. P. Boyle, F. Erben, V. Gülpers, M.T. Hansen, F. Joswig, N.P. Lachini, M. Marshall, A. Portelli, Physical-mass calculation of $\rho(770)$ and $K^*(892)$ resonance parameters via $\pi\pi$ and $K\pi$ scattering amplitudes from lattice QCD (2024). 2406.19193
95. H. Yan, M. Mai, M. Garofalo, U.-G. Meißner, C. Liu, L. Liu, C. Urbach, ω Meson from lattice QCD. Phys. Rev. Lett. **133**, 211906 (2024). [2407.16659]

96. S.M. Dawid, F. Romero-López, S.R. Sharpe, Finite- and infinite-volume study of DDπ scattering. J. High Energy Phys. **01**, 060 (2025). [2409.17059]
97. I. Vujmilovic, S. Collins, L. Leskovec, E. Ortiz-Pacheco, P. Madanagopalan, S. Prelovsek, T_{cc}^{+} via the plane wave approach and including diquark-antidiquark operators. PoS **LATTICE2024**, 112 (2025). [2411.08646]
98. A. Francis, Lattice perspectives on doubly heavy tetraquarks. Prog. Part. Nucl. Phys. **140**, 104143 (2025). [2502.04701]
99. F. Erben, From scattering towards multi-hadron weak decays. PoS **LATTICE2024**, 016 (2025). [2501.19302]
100. S.M. Dawid, Z.T. Draper, A.D. Hanlon, B. Hörz, C. Morningstar, F. Romero-López, S.R. Sharpe, S. Skinner, Two- and three-meson scattering amplitudes with physical quark masses from lattice QCD (2025). 2502.17976
101. N. Lang, D.J. Wilson, D_1 and D_2 resonances in coupled-channel scattering amplitudes from lattice QCD (2025). 2502.04232
102. S.M. Dawid, Z.T. Draper, A.D. Hanlon, B. Hörz, C. Morningstar, F. Romero-López, S.R. Sharpe, S. Skinner, QCD predictions for physical multimeson scattering amplitudes (2025). 2502.14348
103. R.A. Briceño, M.T. Hansen, A. Walker-Loud, Multichannel $1 \to 2$ transition amplitudes in a finite volume. PoS **LATTICE2014**, 095 (2015). [1502.00540]
104. R.A. Briceño, M.T. Hansen, Multichannel $0 \to 2$ and $1 \to 2$ transition amplitudes for arbitrary spin particles in a finite volume. Phys. Rev. D **92**, 074509 (2015). [1502.04314]
105. C.J.D. Lin, G. Martinelli, C.T. Sachrajda, M. Testa, K–> pi pi decays in a finite volume. Nucl. Phys. B **619**, 467 (2001). [hep-lat/0104006]
106. W. Detmold, M.J. Savage, Electroweak matrix elements in the two nucleon sector from lattice QCD. Nucl. Phys. A **743**, 170 (2004). [hep-lat/0403005]
107. H.B. Meyer, Lattice QCD and the timelike pion form factor. Phys. Rev. Lett. **107**, 072002 (2011). [1105.1892]
108. V. Bernard, D. Hoja, U.G. Meissner, A. Rusetsky, Matrix elements of unstable states. J. High Energy Phys. **09**, 023 (2012). [1205.4642]
109. A. Agadjanov, V. Bernard, U.G. Meißner, A. Rusetsky, A framework for the calculation of the $\Delta N\gamma^*$ transition form factors on the lattice. Nucl. Phys. B **886**, 1199 (2014). [1405.3476]
110. R.A. Briceño, M.T. Hansen, A. Walker-Loud, Multichannel $1 \to 2$ transition amplitudes in a finite volume. Phys. Rev. D **91**, 034501 (2015). [1406.5965]
111. X. Feng, S. Aoki, S. Hashimoto, T. Kaneko, Timelike pion form factor in lattice QCD. Phys. Rev. D **91**, 054504 (2015). [1412.6319]
112. R.A. Briceño, J.J. Dudek, R.G. Edwards, C.J. Shultz, C.E. Thomas, D.J. Wilson, The $\pi\pi \to \pi\gamma^\star$ amplitude and the resonant $\rho \to \pi\gamma^\star$ transition from lattice QCD. Phys. Rev. D **93**, 114508 (2016). [1604.03530]
113. K. Polejaeva, A. Rusetsky, Three particles in a finite volume. Eur. Phys. J. A **48**, 67 (2012). [1203.1241]
114. M.T. Hansen, S.R. Sharpe, Relativistic, model-independent, three-particle quantization condition. Phys. Rev. D **90**, 116003 (2014). [1408.5933]
115. M.T. Hansen, S.R. Sharpe, Expressing the three-particle finite-volume spectrum in terms of the three-to-three scattering amplitude. Phys. Rev. D **92**, 114509 (2015). [1504.04248]
116. R.A. Briceño, M.T. Hansen, S.R. Sharpe, Relating the finite-volume spectrum and the two-and-three-particle S matrix for relativistic systems of identical scalar particles. Phys. Rev. D **95**, 074510 (2017). [1701.07465]
117. P. Guo, V. Gasparian, A solvable three-body model in finite volume. Phys. Lett. B **774**, 441 (2017). [1701.00438]
118. H.-W. Hammer, J.-Y. Pang, A. Rusetsky, Three-particle quantization condition in a finite volume: 1. The role of the three-particle force. J. High Energy Phys. **09**, 109 (2017). [1706.07700]

119. H.W. Hammer, J.Y. Pang, A. Rusetsky, Three particle quantization condition in a finite volume: 2. General formalism and the analysis of data. J. High Energy Phys. **10**, 115 (2017). [1707.02176]
120. M. Mai, M. Döring, Three-body unitarity in the finite volume. Eur. Phys. J. A **53**, 240 (2017). [1709.08222]
121. M. Döring, H.W. Hammer, M. Mai, J.Y. Pang, T.A. Rusetsky, J. Wu, Three-body spectrum in a finite volume: the role of cubic symmetry. Phys. Rev. D **97**, 114508 (2018). [1802.03362]
122. R.A. Briceño, M.T. Hansen, S.R. Sharpe, Numerical study of the relativistic three-body quantization condition in the isotropic approximation. Phys. Rev. D **98**, 014506 (2018). [1803.04169]
123. P. Klos, S. König, H.W. Hammer, J.E. Lynn, A. Schwenk, Signatures of few-body resonances in finite volume. Phys. Rev. C **98**, 034004 (2018). [1805.02029]
124. P. Guo, M. Döring, A.P. Szczepaniak, Variational approach to N-body interactions in finite volume. Phys. Rev. D **98**, 094502 (2018). [1810.01261]
125. R.A. Briceño, M.T. Hansen, S.R. Sharpe, Three-particle systems with resonant subprocesses in a finite volume. Phys. Rev. D **99**, 014516 (2019). [1810.01429]
126. T.D. Blanton, F. Romero-López, S.R. Sharpe, Implementing the three-particle quantization condition including higher partial waves. J. High Energy Phys. **03**, 106 (2019). [1901.07095]
127. J.-Y. Pang, J.-J. Wu, H.W. Hammer, U.-G. Meißner, A. Rusetsky, Energy shift of the three-particle system in a finite volume. Phys. Rev. D **99**, 074513 (2019). [1902.01111]
128. F. Romero-López, S.R. Sharpe, T.D. Blanton, R.A. Briceño, M.T. Hansen, Numerical exploration of three relativistic particles in a finite volume including two-particle resonances and bound states. J. High Energy Phys. **10**, 007 (2019). [1908.02411]
129. M.T. Hansen, F. Romero-López, S.R. Sharpe, Generalizing the relativistic quantization condition to include all three-pion isospin channels. J. High Energy Phys. **07**, 047 (2020). [2003.10974]
130. T.D. Blanton, S.R. Sharpe, Alternative derivation of the relativistic three-particle quantization condition. Phys. Rev. D **102**, 054520 (2020). [2007.16188]
131. T.D. Blanton, S.R. Sharpe, Equivalence of relativistic three-particle quantization conditions. Phys. Rev. D **102**, 054515 (2020). [2007.16190]
132. P. Guo, Modeling few-body resonances in finite volume. Phys. Rev. D **102**, 054514 (2020). [2007.12790]
133. F. Romero-López, A. Rusetsky, N. Schlage, C. Urbach, Relativistic N-particle energy shift in finite volume. J. High Energy Phys. **02**, 060 (2021). [2010.11715]
134. J.-Y. Pang, J.-J. Wu, L.-S. Geng, DDK system in finite volume. Phys. Rev. D **102**, 114515 (2020). [2008.13014]
135. T.D. Blanton, S.R. Sharpe, Relativistic three-particle quantization condition for nondegenerate scalars. Phys. Rev. D **103**, 054503 (2021). [2011.05520]
136. F. Müller, T. Yu, A. Rusetsky, Finite-volume energy shift of the three-pion ground state. Phys. Rev. D **103**, 054506 (2021). [2011.14178]
137. F. Müller, A. Rusetsky, On the three-particle analog of the Lellouch-Lüscher formula. J. High Energy Phys. **03**, 152 (2021). [2012.13957]
138. M.T. Hansen, F. Romero-López, S.R. Sharpe, Decay amplitudes to three hadrons from finite-volume matrix elements. J. High Energy Phys. **04**, 113 (2021). [2101.10246]
139. T.D. Blanton, S.R. Sharpe, Three-particle finite-volume formalism for $\pi+\pi+$K+ and related systems. Phys. Rev. D **104**, 034509 (2021). [2105.12094]
140. F. Müller, J.-Y. Pang, A. Rusetsky, J.-J. Wu, Relativistic-invariant formulation of the NREFT three-particle quantization condition. J. High Energy Phys. **02**, 158 (2022). [2110.09351]
141. T.D. Blanton, F. Romero-López, S.R. Sharpe, Implementing the three-particle quantization condition for $\pi^+\pi^+\mathrm{K}^+$ and related systems. J. High Energy Phys. **02**, 098 (2022). [2111.12734]
142. F. Müller, J.-Y. Pang, A. Rusetsky, J.-J. Wu, Three-particle Lellouch-Lüscher formalism in moving frames. J. High Energy Phys. **02**, 214 (2023). [2211.10126]

143. D. Severt, M. Mai, U.-G. Meißner, Particle-dimer approach for the Roper resonance in a finite volume. J. High Energy Phys. **04**, 100 (2023). [2212.02171]
144. A.W. Jackura, R.A. Briceño, M.T. Hansen, Three-pion effects in $K^0 - \bar{K}^0$ mixing. PoS **LATTICE2022**, 062 (2023). [2212.09951]
145. J. Baeza-Ballesteros, J. Bijnens, T. Husek, F. Romero-López, S.R. Sharpe, M. Sjö, The isospin-3 three-particle K-matrix at NLO in ChPT. J. High Energy Phys. **05**, 187 (2023). [2303.13206]
146. Z.T. Draper, M.T. Hansen, F. Romero-López, S.R. Sharpe, Three relativistic neutrons in a finite volume. J. High Energy Phys. **07**, 226 (2023). [2303.10219]
147. R. Bubna, F. Müller, A. Rusetsky, Finite-volume energy shift of the three-nucleon ground state. Phys. Rev. D **108**, 014518 (2023). [2304.13635]
148. R.A. Briceño, M.T. Hansen, Relativistic, model-independent, multichannel $2 \to 2$ transition amplitudes in a finite volume. Phys. Rev. D **94**, 013008 (2016). [1509.08507]
149. A. Baroni, R.A. Briceño, M.T. Hansen, F.G. Ortega-Gama, Form factors of two-hadron states from a covariant finite-volume formalism. Phys. Rev. D **100**, 034511 (2019). [1812.10504]
150. R.A. Briceño, M.T. Hansen, A.W. Jackura, Consistency checks for two-body finite-volume matrix elements: I. Conserved currents and bound states. Phys. Rev. D **100**, 114505 (2019). [1909.10357]
151. R.A. Briceño, M.T. Hansen, A.W. Jackura, Consistency checks for two-body finite-volume matrix elements: II. Perturbative systems. Phys. Rev. D **101**, 094508 (2020). [2002.00023]
152. X. Feng, L.-C. Jin, Z.-Y. Wang, Z. Zhang, Finite-volume formalism in the $2 \xrightarrow{H_I+H_I} 2$ transition: an application to the lattice QCD calculation of double beta decays. Phys. Rev. D **103**, 034508 (2021). [2005.01956]
153. M. Rubin, R. Sugar, G. Tiktopoulos, Dispersion relations for three-particle scattering amplitudes. I. Phys. Rev. **146**, 1130 (1966)
154. M.T. Hansen, S.R. Sharpe, Relativistic three-particle quantization condition: an update. PoS **LATTICE2014**, 088 (2015). [1409.7012]
155. M.T. Hansen, S.R. Sharpe, Perturbative results for two and three particle threshold energies in finite volume. Phys. Rev. D **93**, 014506 (2016). [1509.07929]
156. M.T. Hansen, S.R. Sharpe, Threshold expansion of the three-particle quantization condition. Phys. Rev. D **93**, 096006 (2016). [1602.00324]
157. M.T. Hansen, S.R. Sharpe, Applying the relativistic quantization condition to a three-particle bound state in a periodic box. Phys. Rev. D **95**, 034501 (2017). [1609.04317]
158. M.T. Hansen, S.R. Sharpe, Lattice QCD and three-particle decays of resonances. Ann. Rev. Nucl. Part. Sci. **69**, 65 (2019). [1901.00483]

Schwinger–Keldysh Formalism

3

Felix Haehl and Mukund Rangamani

Abstract

These notes provide an overview of real-time techniques in quantum field theories and holography. We outline the general rationale and principles underlying the Schwinger-Keldysh formalism and its out-of-time-order generalizations. To illustrate its broad utility we frame our discussion in the language of open quantum dynamics. As a specific application we describe how such real-time observables help in understanding scrambling dynamics. We also describe the holographic prescription for computing thermal Schwinger-Keldysh observables.

3.1 Invitation to the Schwinger–Keldysh Path Integral

Mukund Rangamani

The Schwinger-Keldysh formalism is broadly aimed at facilitating the computation of real-time observables in quantum systems. In the following, we give a broad physical motivation for it, before diving into the details. The discussion below largely follows the presentation in [1] and is also reviewed in [2].

F. Haehl
School of Mathematical Sciences, University of Southampton, Southampton, UK
e-mail: F.M.Haehl@soton.ac.uk

M. Rangamani (✉)
Center for Quantum Mathematics and Physics (QMAP), Department of Physics and Astronomy, University of California, Davis, CA, USA
e-mail: mrangamani@ucdavis.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

N. Arkani-Hamed et al. (eds.), *Records from the S-Matrix Marathon*,
Lecture Notes in Physics 1041, https://doi.org/10.1007/978-3-031-90352-6_3

Let us consider computing in a QFT the two-point Green's function for some (generically complex) Heisenberg operator $\mathcal{O}(x)$ in some pure state

$$G(x,x') = -i\,\langle\Omega|\mathcal{T}\,\mathcal{O}(x)\,\mathcal{O}(x')|\Omega\rangle\ , \tag{3.1.1}$$

where $\mathcal{T}$ is the standard time ordering, and $x = (t, \mathbf{x})$ is the spacetime coordinate. We will distinguish the temporal direction (coordinatized by t) when necessary. The state $|\Omega\rangle$ is the ground state of the full interacting theory. In perturbation theory, working in the interaction picture, the non-trivial part of the evolution operator

$$U(t_0,t) = \mathcal{T}\,\exp\left[-i\int_{t_0}^{t}\mathrm{d}t'\,H_{\rm int}(t')\right]\ , \tag{3.1.2}$$

defines temporal evolution of the interaction picture states. Using this expansion, one then finds an expression for the two-point Green's function:

$$G(x,x') = -i\,\langle 0|S^\dagger\,\mathcal{T}\,\mathcal{O}(x)\,\mathcal{O}^\dagger(x')\,S|0\rangle = -i\,\frac{\langle 0|\mathcal{T}[S\,\mathcal{O}(x)\,\mathcal{O}^\dagger(x')]|0\rangle}{\langle 0|S|0\rangle}\ . \tag{3.1.3}$$

We arrive at this equation by introducing the S-matrix $S \equiv U(-\infty,\infty)$ and obtaining the interacting ground state by evolving the non-interacting ground state $|0\rangle$ of the free Hamiltonian H_0. We usually use the final expression in (3.1.3) as the starting point for perturbation theory.

In writing the second equality, we have expressed the instantaneous late time ground state in terms of the early time state, assuming an adiabatic evolution of the system expressed as a property of the S-matrix. Namely, the phase picked up by acting on the final state with $S^\dagger$ is the same as the one accumulated during the evolution, i.e., $\langle 0|\,S^\dagger = \langle 0|\,\mathrm{e}^{-i\alpha}$ and $\langle 0|S|0\rangle = \mathrm{e}^{i\alpha}$, for some phase α. Thus, we assume that the physical content of the ground state remains unchanged during the evolution, up to a phase rotation. This fails in non-equilibrium situations, where adiabatic evolution is not justified.

The Schwinger–Keldysh (SK) formalism deals with non-equilibrium dynamics by only ever making reference to the initial state,[1] which may be taken without loss of generality to be an equilibrium configuration, the instantaneous vacuum state of H_0 at $t = -\infty$. As we evolve the system, we want to ensure that we make no assumption about where (or what state) it would end up at late times. To this end, rather than continuing to evolve forward, we should revert, after allowing the interactions to influence the system, to the initial state. In a path integral, this can be done systematically by introducing a SK-evolution operator, which evolves the system in a complex time contour. Let $\mathcal{C}$ be a contour in the complex time plane that starts out at $t = -\infty + i\varepsilon$, follows the real axis, and then retraces its trajectory

[1] Hence, the Schwinger–Keldysh formalism is sometimes also referred to as *in-in formalism*.

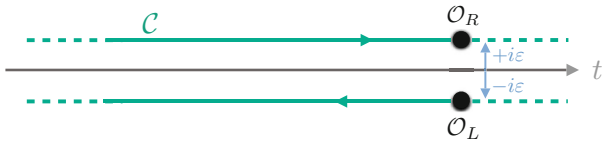


Fig. 3.1 An illustration of the generic Schwinger–Keldysh complex time contour. Every operator $\mathcal{O}$ in the original theory has two representations in the Schwinger–Keldysh path integral, viz., $\mathcal{O}_R$ and $\mathcal{O}_L$. The labels help distinguish which part of the contour the operator is inserted on. There is a natural contour ordering prescription wherein right operators are time-ordered and left operators are anti-time ordered

back with a small imaginary displacement by $-2i\varepsilon$, cf. Fig. 3.1. We have chosen to orient the contour so that the direction of traversal is clockwise (about the origin, say). We will also find it useful to label the forward leg of the contour as the right (R) part and the backward leg the left (L) part.

Given such a contour, we can work with operators which live in this complexified domain, and define the Schwinger–Keldysh S-matrix by working with contour-ordering prescription, to wit,

$$U_{\mathcal{C}} \equiv \mathcal{T}_{\mathcal{C}} \exp\left[-i \int_{\mathcal{C}} \mathrm{d}t' \, H_{\mathrm{int}}(t')\right] . \tag{3.1.4}$$

There is a sensible time-ordering prescription inherited from this contour-ordering.

It is often, however, useful not to work with a single contour, but rather, work with fields and operators labeled by which part of the contour they appear on. This perspective is best served by doubling of the degrees of freedom and using it to obtain physical results. We have left and right fields indexed by their position on the Schwinger–Keldysh contour $\mathcal{C}$. Furthermore, as illustrated, the operators on the right/forward leg are time-ordered, those on the left/backward leg are anti-time-ordered, and the right operators precede those on the left leg of the contour.

If we have complete knowledge of the state (which we can take in general to be a density matrix) of the system at some finite time t_0, then we do not need to follow the contour all the way from $t = -\infty$ to $+\infty$ and back. It suffices to focus solely on the part of the contour from t_0 to $\max(t, t')$, which corresponds to the future-most operator insertion, before retracing back to the initial configuration, cf. Fig. 3.2. Intuitively, all this is saying is that the knowledge of the density matrix can be treated as initial conditions for the subsequent evolution. Furthermore, for finite time computations, details of how the system evolves to the future of all operator insertions (or measurements) are inessential.

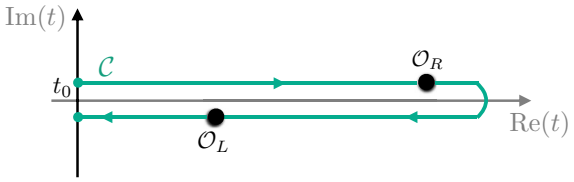


Fig. 3.2 SK time contour in the case where the initial state at time t_0 is known and the latest operator insertion happens at time t. The indicated operator insertions correspond to a real-time correlator $G_<(x, x')$

3.1.1 The Schwinger–Keldysh Path Integral

Based on the discussion above, we can motivate the general Schwinger–Keldysh path integral,

$$\mathcal{Z}_{\rm SK}[J_R, J_L] = {\rm Tr}\Big[U[J_R]\,\rho_0\,U^\dagger[J_L]\Big]\,, \tag{3.1.5}$$

where R and L refer to the evolution of the ket and bra part of the density matrix. This computes SK-path ordered correlators, which consist of a product of time-ordered operator insertions followed by a product of anti-time-ordered ones. We denote an operator inserted on contour segment $I \in \{R, L\}$ as $\mathcal{O}_I$. A typical SK correlator is of the form[2]

$$\begin{aligned}&\big\langle \mathcal{T}_{\mathcal{C}}\,\mathcal{O}_R(t_1)\cdots\mathcal{O}_R(t_k)\mathcal{O}_L(t_{k+1})\cdots\mathcal{O}_L(t_n)\big\rangle \\ &\qquad = {\rm Tr}\big[\big(\bar{\mathcal{T}}\,\mathcal{O}(t_n)\cdots\mathcal{O}(t_{k+1})\big)\big(\mathcal{T}\,\mathcal{O}(t_k)\cdots\mathcal{O}(t_1)\big)\,\rho_0\big]\,,\end{aligned} \tag{3.1.6}$$

where $\mathcal{T}_{\mathcal{C}}$, $\mathcal{T}$, $\bar{\mathcal{T}}$ denote contour ordering, time ordering, and anti-time ordering, respectively. The R operators, which correspond to the ket part of ρ, are time-ordered, while the L operators, which correspond to the bra part, are anti-time-ordered.

As presaged, the original motivation for the Schwinger–Keldysh formalism is to describe systems interacting with an environment. In this context, it is natural to work with variables that evolve classically, and those that are subject to quantum fluctuations from the environment. This is achieved by switching to the average/difference basis, where difference fields (quantum) parametrize fluctuations:

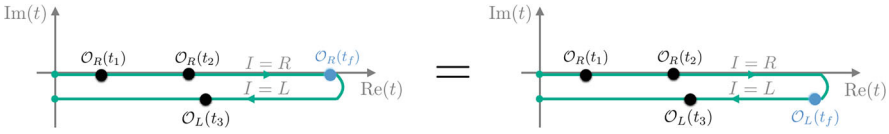
$$\mathcal{O}_{\rm av} = \frac{\mathcal{O}_R + \mathcal{O}_L}{2}\,, \qquad \mathcal{O}_{\rm dif} = \mathcal{O}_R - \mathcal{O}_L\,. \tag{3.1.7}$$

[2] Of course, the operator insertions do not all have to be the same. We have adopted a succinct notation, suppressing additional labels which would serve to distinguish operators.

In this basis, the so-called *largest time equation* is

$$\langle \mathcal{T}_{\mathcal{C}}\, \mathcal{O}_{\rm dif}(t_f)\, \mathcal{O}_{I_1}(t_1) \cdots \mathcal{O}_{I_n}(t_n) \rangle = 0 \qquad (t_f > t_1, \ldots, t_n)\,, \tag{3.1.8}$$

where $\mathcal{O}_{I_i}$ are arbitrary operator combinations on the contours ($I_i = R, L$ or $I_i = \{\text{av}, \text{dif}\}$ label the contour segments), and $\mathcal{T}_{\mathcal{C}}$ denotes Schwinger–Keldysh contour ordering. These constraints follow from unitarity: $U^\dagger[J]U[J] = 1$ when $J_R = J_L \equiv J$ since $\mathcal{O}_{\rm av}$ is sourced by this J. Pictorially:



Generally, the coupling between sources and operators is $J_R\, \mathcal{O}_R - J_L\, \mathcal{O}_L = J_{\rm av}\, \mathcal{O}_{\rm dif} + J_{\rm dif}\, \mathcal{O}_{\rm av}$.

Largest Time Sum Rule The natural objects on the Schwinger–Keldysh contour are the contour-ordered correlators. Consider first the Schwinger–Keldysh two-point function matrix $\mathbb{G}_{IJ}(x, x') = \langle \mathcal{T}_{\mathcal{C}}\, \mathcal{O}_I(x) \mathcal{O}_J^\dagger(x') \rangle$, where $I, J \in \{R, L\}$. These can be assembled into a matrix

$$\mathbb{G}(x, x') = \begin{pmatrix} iG_F(x, x') & iG_>(x, x') \\ iG_<(x, x') & iG_{\bar F}(x, x') \end{pmatrix} = \begin{pmatrix} \langle \mathcal{T}\, \mathcal{O}(x) \mathcal{O}^\dagger(x') \rangle & \langle \mathcal{O}^\dagger(x') \mathcal{O}(x) \rangle \\ \langle \mathcal{O}(x) \mathcal{O}^\dagger(x') \rangle & \langle \overline{\mathcal{T}}\, \mathcal{O}(x) \mathcal{O}^\dagger(x') \rangle \end{pmatrix}. \tag{3.1.9}$$

In our earlier language,

$$\begin{aligned} G_F(x, x') &= -i\, \langle \Omega | \mathcal{T}\, \mathcal{O}(x)\, \mathcal{O}(x') | \Omega \rangle\,, \\ G_{\overline{F}}(x, x') &= -i\, \langle \Omega | \overline{\mathcal{T}}\, \mathcal{O}(x)\, \mathcal{O}(x') | \Omega \rangle\,, \\ G_<(x, x') &= -i\, \langle \Omega | \mathcal{O}(x')\, \mathcal{O}(x) | \Omega \rangle\,, \\ G_>(x, x') &= -i\, \langle \Omega | \mathcal{O}(x)\, \mathcal{O}(x') | \Omega \rangle\,. \end{aligned} \tag{3.1.10}$$

These satisfy the sum rule

$$G_F + G_{\bar F} - G_> - G_< = 0\,. \tag{3.1.11}$$

This sum rule encapsulates once again the unitarity and causality constraints. It is the simplest example of a more general statement. Consider all contour-ordered n-point functions; one can readily check the following identity:[3]

[3] Here, $I_i = 1, 2$ correspond to labels R, L.

$$\sum_{I_1,\ldots,I_n\in\{1,2\}} (-1)^{I_1+\ldots+I_n} \langle \mathcal{T}_{\mathcal{C}}\, \mathcal{O}_{I_1}(x_1)\cdots\mathcal{O}_{I_n}(x_n)\rangle = 0\,. \tag{3.1.12}$$

The cancellation in this sum occurs pairwise: assume without loss of generality that $t_1 > t_2, \ldots, t_n$. Then for every choice of $I_2, \ldots, I_n$ there are two different choices for I_1. The difference between these two choices is zero thanks to the largest time Eq. (3.1.8). We will refer to (3.1.12) as the *largest time sum rule*.

Given the SK generating function (3.1.5), one can compute all contour-ordered correlators by varying with respect to the sources J_R and J_L. Note that this is a highly redundant picture since there are a priori 2^n contour ordered n-point functions, which is greater than the number of Wightman functions (which number $n!$). Some of these redundancies are accounted for by the largest time equations. In fact, a useful way to encode this is to invoke a BRST symmetry of the Schwinger-Keldysh construction, see the Appendix for a brief sketch of this idea. Before we turn to finding a simple algorithm to map the contour-ordered observables to physical ones, it is useful to consider a generalization.

3.1.2 The OTO Contours

Consider a generic n-point function of Heisenberg operators, $\mathcal{O}^{(i)}(t_i)$, that is, the Wightman correlation functions $\left\langle \mathcal{O}^{(1)}(t_1)\,\mathcal{O}^{(2)}(t_2)\,\cdots\,\mathcal{O}^{(n)}(t_n)\right\rangle$, without prescribed temporal ordering. Writing out this correlation function in terms of Schrödinger operators, $\mathcal{O}^{(i)}(t_0)$, using $\mathcal{O}(t) = U(t_0,t)^\dagger\,\mathcal{O}(t_0)\,U(t_0,t)$, we obtain

$$\begin{aligned} G(t_1,\cdots,t_n) &\equiv \left\langle \mathcal{O}^{(1)}(t_1)\,\mathcal{O}^{(2)}(t_2)\,\cdots\,\mathcal{O}^{(n)}(t_n)\right\rangle \\ &= \left\langle U^\dagger(t_0,t_1)\,\mathcal{O}^{(1)}(t_0)\,U(t_0,t_1)\,\cdots U^\dagger(t_0,t_n)\,\mathcal{O}^{(n)}(t_0)\,U(t_0,t_n)\right\rangle. \end{aligned} \tag{3.1.13}$$

The temporal evolution of the system between the operator insertions involves a series of forward and backward evolutions by $U(t_0,t_i)$ and $U^\dagger(t_0,t_j)$, respectively. This is an inevitable consequence of the lack of any temporal ordering. One can represent such an evolution by a path integral contour, which involves a series of temporal switchbacks; see Fig. 3.3.

We will refer to such contours as the *timefold* or *out-of-time-order* (OTO) contours. It is useful to decorate the latter with a depth label, k, with a k-OTO contour specifying a contour, with k forward and k backward evolutions. The Schwinger–Keldysh contour is then a special case, being the 1-OTO contour.

The generating function for the k-OTO contour can be computed with sources J for the operators $\mathcal{O}$, and encoded in the evolution operator $U[J]$

$$U[J] = \mathcal{T}\,\exp\left[-i\int_{t_i}^{t} dt\, H[J]\right], \qquad (U[J])^\dagger = \bar{\mathcal{T}}\,\exp\left[i\int_{t_i}^{t} dt\, H[J]\right]. \tag{3.1.14}$$

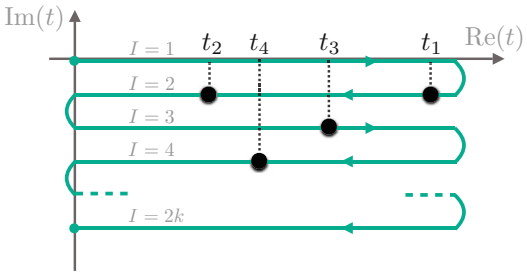


Fig. 3.3 The k-OTO contour, computing the out-of-time-ordered correlation functions encoded in the generating functional (3.1.15). As in Figs. 3.1 and 3.2 the physical time runs from left to right, and the vertical excursions are infinitesimal ($i\varepsilon$). We also show insertion points required to compute the correlator with temporal ordering $t_1 > t_2$, $t_2 < t_3$, and $t_3 > t_4$

The k-OTO generating function is then defined to be

$$\mathcal{Z}_{k-\text{oto}}[J_I] = \text{Tr}[\cdots U[J_3](U[J_2])^\dagger U[J_1]\ \rho_i\ (U[J_{2k}])^\dagger U[J_{2k-1}](U[J_{2k-2}])^\dagger \cdots]\,. \tag{3.1.15}$$

The contour prescription with the decorated labels $I = 1, \ldots, 2k$ is useful, but highly redundant. There are several useful bases that are adapted to obtain physical results more directly [3]. We describe them below, with fixed time order, $t_1 > t_2 > \cdots > t_n$, for definiteness:

- Wightman basis: This comprises the $n!$ physical correlators

$$G_\sigma(t_1, t_2, \cdots, t_n) = \left\langle \mathcal{O}^{(\sigma_1)}\, \mathcal{O}^{(\sigma_2)}\, \cdots\, \mathcal{O}^{(\sigma_n)} \right\rangle, \qquad \sigma \in S_n\,, \tag{3.1.16}$$

  where S_n denotes the group of permutations of n objects.
- Nested correlators: Another useful set uses nested commutators and anti-commutators of the n operators $\mathcal{O}^{(i)}$. These are constructed in terms of the elementary building blocks, which are commutators $[\cdot, \cdot]$ and anti-commutators $\{\cdot, \cdot\}$ of the operators. We can consider nesting a sequence of graded commutator, anti-commutators; for example,

$$[\{[\mathcal{O}^{(1)}, \mathcal{O}^{(2)}], \mathcal{O}^{(3)}\}, \cdots]\,, \tag{3.1.17}$$

  which illustrates the general idea. This set is spanned by $2^{n-2}n!$ correlators. These objects, together with appropriate time-ordering step functions, form the basis of time-ordered response functions [1] (also see [2]). This statement should be familiar for 2-point functions, since the complete information of the propagator is contained in the commutator and anti-commutator. The reason for their importance can be traced to the fact that Lorentzian causal ordering ensures

that the commutator of operators will vanish when the insertions are spacelike separated.

- The LR correlators: The space of correlation functions derived from the k-OTO contour forms another basis. There are a total of $(2k)^n$ correlation functions, since $\mathcal{O}(t_i)$ can be inserted in any one of the $2k$ legs of the contour. This vast overdetermination can be collapsed to something simpler, for instance, by switching off or aligning some sources, we can collapse some timefolds using unitarity.
- The Av-Dif correlators: A closely related basis involves rotating the LR-basis into the average-difference operator basis. This is done by a natural extension of the Keldysh basis used in the usual Schwinger–Keldysh formalism $\mathcal{O}_{\text{av}(\ell)} = \frac{1}{2}\left(\mathcal{O}_{2\ell-1} + \mathcal{O}_{2\ell}\right)$ and $\mathcal{O}_{\text{dif}(\ell)} = \mathcal{O}_{2\ell-1} - \mathcal{O}_{2\ell}$, where $\ell = 1, \ldots, k$.

Note that in situations where the state is analytic, we may use an $i\varepsilon$ prescription to recover the Wightman function from the Euclidean Schwinger functions, via

$$G_\sigma(t_1, t_2, \cdots, t_n) = \lim_{\varepsilon_i \to 0} G_E(\tau_1, \tau_2, \cdots, \tau_n)\Big|_{\tau_i = i\, t_i + \varepsilon_i} \qquad (\varepsilon_{\sigma_1} > \ldots > \varepsilon_{\sigma_n})\,. \tag{3.1.18}$$

This can be implemented directly in 2d CFTs thanks to the convergence of the OPE. A nice discussion of higher-dimensional CFTs can be found in [4].

Let us record some general results about these OTO contours:

- The computation of n-point Wightman functions can be done on a k-OTO contour with $k \leqslant \lfloor \frac{n+1}{2} \rfloor$. This can be seen by noting that the worst-case scenario is an ordering in which we first act with $\mathcal{O}(t_1)$, then by $\mathcal{O}(t_n)$, followed by $\mathcal{O}(t_2)$,and then acting with $\mathcal{O}(t_{n-1})$ and so on, involving switching back between forward and backward evolutions.
- One can trivially embed an n-point function on a $k > \lfloor \frac{n+1}{2} \rfloor$-OTO contour. This will involve some redundancies, but it is useful to derive relations between correlators.
- For completeness, let us record that all two-point functions can be computed on a SK/1-OTO contour. Of the 6 three-point functions, 4 are computed on the 1-OTO contour, but two others $\langle \mathcal{O}(t_1)\, \mathcal{O}(t_3)\, \mathcal{O}(t_2) \rangle$, and $\langle \mathcal{O}(t_2)\, \mathcal{O}(t_3)\, \mathcal{O}(t_1) \rangle$ require the use of a 2-OTO contour. Similarly, only 8 of the 24 four-point functions can be computed on the 1-OTO contour; the remaining 16 require a 2-OTO contour.

The redundancies on the k-OTO contour can be understood pictorially by thinking of the contour as a wireframe of an abacus, with the operators as beads. The operators can slide up and down to occupy their temporal spot on any of the legs, as long as other operators do not obstruct them. The basic reason for this is again the analog of the largest time equation, which originates from the normalization of the time-ordering step function. We define

$$\Theta_{abcd\cdots} = \Theta_{ab}\, \Theta_{bc}\, \Theta_{cd} \cdots, \qquad \Theta_{ab} = 1\,, \quad \text{if } t_a > t_b\,. \tag{3.1.19}$$

From here it follows that,

$$\sum_{\sigma \in S_n} \Theta_{\sigma_1 \sigma_2 \cdots \sigma_n} = 1 \,. \tag{3.1.20}$$

As noted above, this implies that difference operators cannot be future most. This observation can be used to give a simple mapping between the different bases described above using the *Keldysh bracket*. We will illustrate it below for the 1-OTO case, and refer to [3] for the k-OTO generalization.

The Keldysh bracket is defined as an operation that takes as input a physical single copy (non-Schwinger–Keldysh) operator as its first entry, a Schwinger-Keldysh operator as the second entry, and outputs a product of the single copy operators as the result. Letting I be the identity operator, one defines

$$\begin{aligned} (A, B_R)_{\rm SK} &\equiv A\, B \,, & (A, B_L)_{\rm SK} &\equiv B\, A \,, \\ (A, B_{\rm av})_{\rm SK} &\equiv \{A, B\} \,, & (A, B_{\rm dif})_{\rm SK} &\equiv [A, B] \,, \\ (I, B_{\rm av})_{\rm SK} &\equiv B \,, & (I, B_{\rm dif})_{\rm SK} &\equiv 0 \,, \end{aligned} \tag{3.1.21}$$

A Schwinger–Keldysh correlator can be expressed as a nested Keldysh bracket acting on the identity operator and then applying Schwinger–Keldysh time-ordering. The time-ordering is a particular choice of the step functions. One simply sums over all possible orderings of operators inside the nested Keldysh brackets and dresses each of them with the appropriate causal step function. Specifically,

$$\begin{aligned} &\langle \mathcal{T}_{SK}\, \mathcal{O}^{(1)}_{I_1} \mathcal{O}^{(2)}_{I_2} \dots \mathcal{O}^{(p)}_{I_p} \rangle \\ &\quad = \sum_{\substack{\text{time} \\ \text{orderings}}} \Theta_{\sigma_1 \sigma_2 \cdots \sigma_p}\, \langle (\cdots ((I, \mathcal{O}^{(\sigma_1)})_{\rm SK}, \mathcal{O}^{(\sigma_2)})_{\rm SK}, \cdots, \mathcal{O}^{(\sigma_p)})_{\rm SK} \rangle \,, \end{aligned} \tag{3.1.22}$$

where $\sigma_1 \sigma_2 \cdots \sigma_p$ is a permutation of the p indices. This expression can be used to bring any correlation function to a standard form involving commutators and anti-commutators. For instance, one can check that

$$\begin{aligned} \langle \mathcal{T}_{SK}\, A_{\rm av}\, B_{\rm av} \rangle &= \Theta_{AB}\, \left\langle ((I, A_{\rm av})_{\rm SK}, B_{\rm av})_{\rm SK} \right\rangle + \Theta_{BA}\, \left\langle ((I, B_{\rm av})_{\rm SK}, A_{\rm av})_{\rm SK} \right\rangle \\ &= \Theta_{AB}\, \left\langle (A, B_{\rm av})_{\rm SK} \right\rangle + \Theta_{BA}\, \left\langle (B, A_{\rm av})_{\rm SK} \right\rangle = \langle \{A, B\} \rangle \,. \end{aligned} \tag{3.1.23}$$

Similarly, we may further deduce that

$$\begin{aligned} \langle \mathcal{T}_{SK}\, A_{\rm av}\, B_{\rm dif} \rangle &= \Theta_{AB}\, \langle [A, B] \rangle \\ \langle \mathcal{T}_{SK}\, A_{\rm av}\, B_{\rm dif}\, C_{\rm dif} \rangle &= \Theta_{ABC}\, \langle [[A, B], C] \rangle + \Theta_{CAB}\, \langle [[C, A], B] \rangle \,. \end{aligned} \tag{3.1.24}$$

Let us take stock: the space of n-point correlation functions is, in general, spanned by $n!$ distinct Wightman functions. These can be represented on OTO contours in many equivalent ways. To classify them, we can denote as $g_{n,q}$ the number of *proper* q-OTO n-point functions, i.e., those which require at least a q-OTO path integral to be representable. A proper q-OTO n-point function can be represented on a k-OTO contour (with $k \geqslant q$) in some number of ways, $h^{(q)}_{n,k}$. This leads to a decomposition of the number of n-point functions into equivalence classes under unitarity rules:

$$n! = \sum_{q=1}^{\lfloor \frac{n+1}{2} \rfloor} g_{n,q} \,, \qquad (2k)^n = \sum_{q=1}^{\lfloor \frac{n+1}{2} \rfloor} g_{n,q}\, h^{(q)}_{n,k} \,, \tag{3.1.25}$$

where we recall that $(2k)^n$ is the (vastly redundant) total number of n-point functions representable on the k-OTO contour: each of the n operators can be inserted on any of the $2k$ contour segments. The unitarity rules translate into a combinatorial counting problem, which allows us to compute the integers $g_{n,q}$ and $h^{(q)}_{n,k}$ recursively. We refer to [3] for details and results.

3.2 Schwinger–Keldysh and Cutkosky Cutting Rules

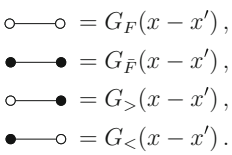
Mukund Rangamani

We will now briefly recall the cutting rules and perturbative unitarity, showing that this formalism is isomorphic to the Schwinger–Keldysh approach described above, cf. [5] for the original discussion.

For concreteness, consider Feynman diagrams built out of propagators

$$\begin{aligned} G_F(x) &= \Theta(x^0) G_>(x) + \Theta(-x^0) G_<(x) \,, \\ G_{\gtrless}(x) &= \int \frac{\mathrm{d}^4 k}{(2\pi)^4}\, \delta(k^2 - m^2)\, \Theta(\pm k^0)\, \mathrm{e}^{-ikx} \,. \end{aligned} \tag{3.2.1}$$

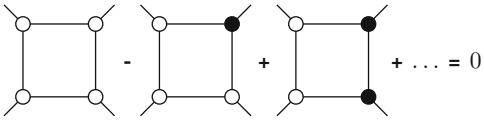
Here $G_{\gtrless}(x)$ are the two Wightman functions and refer to the positive and negative frequency parts of the Feynman propagator. Decomposing Feynman diagrams into positive and negative frequency components leads to an efficient way of tracking causality constraints. One introduces a notation where every vertex in a diagram is either circled (black) or uncircled (white) with the following rules for propagators connecting the various vertex combinations:



(3.2.2)

In addition, every circled vertex receives a minus sign. Physically, positive energy is always transferred from uncircled to circled vertices. Clearly, the effective doubling of degrees of freedom (by means of doubling every vertex) is very much analogous to the Schwinger–Keldysh approach. Let us discuss this in more detail.

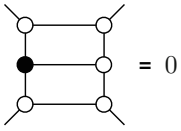
The largest time sum rule is now the statement that the sum over all circlings of a given diagram vanishes. For example:



(3.2.3)

The cancellation again happens in pairs: any given diagram cancels against its 'largest time partner' diagram, i.e., the diagram obtained by changing the circle on the future-most vertex.

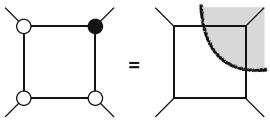
Note that generally many diagrams vanish because of the condition that energy can only flow into circled vertices. For instance, an isolated circled vertex that is not connected to any external legs must give a vanishing contribution as energy cannot flow to it from anywhere:



(3.2.4)

This constraint is obvious in momentum space, where the diagram is proportional to a vanishing product of $\Theta(\pm k^0_{ij})$. This makes it less transparent in Schwinger–Keldysh formalism.

We can obtain a more transparent statement by considering again the sum over all circlings in frequency space. Every choice of circlings can be associated with a cut that separates the diagram into a shaded (circled) and sunny (uncircled) part. Energy flows from the unshaded region into the shaded region, and both regions must be connected to external legs in a way that is consistent with this (otherwise the diagram vanishes). For instance:



(3.2.5)

with the understanding that we put all cut internal propagators on-shell,

$$\frac{1}{k^2 - m^2} \longrightarrow \delta(k^2 - m^2)\,, \tag{3.2.6}$$

thus turning the cut diagram into a product of amplitudes. The *Cutkosky cutting rule* is precisely the largest time sum rule written as a sum over cuts:

$$F(k_1, \ldots, k_n) + \bar{F}(k_1, \ldots, k_n) = -\sum_{\text{cuts}} F_{\text{cut}}(k_1, \ldots, k_n)\,, \tag{3.2.7}$$

where F is a (momentum space) diagram with all vertices uncircled, $\bar{F}$ is the corresponding diagram with all vertices circled, and F_{cut} are the various cut versions of the same diagram.

We conclude this discussion by making the connection with the S-matrix manifest. Recall that unitarity of $S = 1 + iT$ implies

$$T - T^\dagger = iT^\dagger T\,. \tag{3.2.8}$$

The cutting rule (3.2.7) looks precisely like the contribution of a given diagram F to this unitarity condition, where T accounts for sunny amplitudes and $T^\dagger$ for shaded ones. This is a little too fast, but it does lead us in the right direction: in particular, in order for (3.2.7) to really encode (3.2.8), we need that the Lagrangian (which defines the Feynman rules for F) satisfies $\mathcal{L} = \mathcal{L}^\dagger$. This will then ensure that the Feynman rules used to build a diagram $\bar{F}$ do indeed correspond to those associated with $S^\dagger$ (or $T^\dagger$).

3.3 The Open Quantum System Paradigm

Mukund Rangamani

As described above, the Schwinger–Keldysh formalism and its k-OTO generalization provide a framework for computing real-time observables in QFT. One concrete application of these ideas is in the context of non-equilibrium dynamics. We will frame this application in terms of the open quantum system paradigm, wherein some quantum degrees of freedom are traced over.

Consider a QFT with some degrees of freedom encoded in local fields $\Psi(x, t)$, which is the physical system of interest. We will imagine this system coupled to an environment, which itself is another field theory, perhaps with many degrees of freedom X_i. The unitary microscopic theory is of the form:

$$S[\Psi, X_i] = S_s[\Psi] + S_e[X_i] + S_{s-e}[\Psi, X_i]\,. \tag{3.3.1}$$

The combined system and environment is initialized in some initial state, which could even be a factorized state $\rho_s \otimes \rho_e$. This factorized form will fail to hold under evolution, which is generated by the joint system-environment Hamiltonian obtained from the above action.

If we integrate out the environment degrees of freedom X_i, we will end up with a density operator for the system. This system density operator will itself undergo a non-unitary evolution. The paradigm for such open system dynamics was laid

out in [6]. A key observation in this work was that the natural way to describe the evolution of the system is in terms of a Schwinger–Keldysh path integral, with a non-trivial interaction between the L and R degrees of freedom, which is referred to as the *influence function*.

$$\begin{aligned}&\int [D\Psi_R][D\Psi_L] \int [DX_{i,R}][DX_{i,L}]\, \mathrm{e}^{i\,S[\Psi_R,X_{i,R}]-i\,S[\Psi_L,X_{i,L}]} \\ &\quad = \int [D\Psi_R][D\Psi_L]\, \mathrm{e}^{i\,S[\Psi_R]-i\,S[\Psi_L]+i\,S_{IF}[\Psi_R,\Psi_L]}\,.\end{aligned} \tag{3.3.2}$$

The influence functional is clearly induced onto the system owing to the coupling with the environment.

This paradigm has broad relevance, ranging from the analysis of environment-induced effects, cosmology, and thermal systems. An approach that is usually taken in these contexts is to assume that the environment's dynamics is suitably Markovian. By this, one means that the system-environment interactions are such that any information exchanged between them is rapidly dissipated or "forgotten" by the environment. In other words, the environment has no long-term memory. This implies that the system density matrix will evolve by a trace-preserving positive map (a quantum channel). In quantum mechanics, one can, with these assumptions, write down a master equation for the evolution, generalizing the standard unitary evolution. This takes the form

$$i\,\frac{\mathrm{d}\rho}{\mathrm{d}t} = [H,\rho] + i\sum_{a,b}\Gamma_{ab}\left(L_b\,\rho\,L_a^\dagger - \frac{1}{2}\,L_a^\dagger\,L_b\,\rho - \frac{1}{2}\,\rho\,L_a^\dagger\,L_b\right). \tag{3.3.3}$$

Here, L_a and L_b are positive operators indexed by a, b, and $\Gamma_{a,b}$ are effective couplings. The evolution can be seen to be trace-preserving.

In general, the Markovianity assumption can fail, especially in situations where the environment has some long-lived modes dominating its late-time dynamics. Integrating these out will clearly lead to some non-local effects. In such situations, the influence functional is more directly useful.

Note that in the absence of the influence functional, the dynamics of the system is factorized between the ket and bra (R and L). All non-trivial correlations in this case are all captured by the initial density matrix. The factorization fails when we have influence functionals. However, these cannot be completely arbitrary, since that would violate the microscopic unitarity (before integrating out the environment). The influence functional is constrained by the largest time equation, which leads to a set of Lindblad conditions in the case of Markovian dynamics; cf. [7] for an analysis in the context of renormalization of an open QFT.

An arena that is well suited to develop the structural aspects of open effective field theories (open EFTs) is holography [8, 9]. We imagine the environment to be a large N holographic field theory, which is strongly coupled. A class of such theories has a simple dual description in terms of gravitational dynamics in an asymptotically

AdS spacetime. The prototype example of this is the four-dimensional $\mathcal{N} = 4$ SYM theory with gauge group SU(N). In the large N limit, the theory has a planar diagram expansion with 't Hooft coupling $\lambda = g_{YM}^2 N$. The dual gravitational theory lives on an AdS spacetime with curvature length scale ℓ_{AdS} which is hierarchically larger than the string and Planck scales, ℓ_s and ℓ_P, respectively. The parameters are related through,

$$\lambda \sim \left(\frac{\ell_{\text{AdS}}}{\ell_s}\right)^4 , \qquad N^2 \sim \left(\frac{\ell_{\text{AdS}}}{\ell_P}\right)^4 . \tag{3.3.4}$$

We will describe below the general setup (cf. Sect. 3.5) for computing Schwinger–Keldysh correlators in such holographic field theories. These enter as the effective couplings of the open quantum system coupled to a holographic environment.

3.4 Thermal States and Advanced/Retarded Basis

Mukund Rangamani

An interesting and important class of examples is the behavior of systems coupled to thermal environments. To understand these, we can specialize to considering the initial state to be a thermal density operator

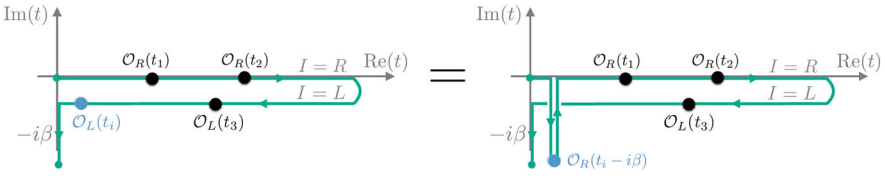
$$\rho_\beta = \mathrm{e}^{-\beta H} . \tag{3.4.1}$$

3.4.1 The KMS Condition

Since the thermal density operator involves an evolution in the Euclidean time direction (by an amount β), we can define a useful notion of conjugating operators through the density operator, and inserting them at imaginary time values:

$$\mathcal{O}(t - i\beta) = \rho_\beta^{-1}\, \mathcal{O}(t)\, \rho_\beta . \tag{3.4.2}$$

If we view the conjugation operation as passing an operator through the density matrix, then the motion is counterclockwise along the contour. In fact, viewing the density operator itself as an evolution in imaginary time, the conjugation pictorially can be visualized by sliding the smallest-time operator across the initial state:



As before, we define the one-sided thermal SK ('1-OTO') path integral as

$$\mathcal{Z}^{\beta}_{\rm SK}[J_R, J_L] = {\rm Tr}(U[J_R]\,\rho_\beta\,(U[J_L])^\dagger)\,. \tag{3.4.3}$$

This path integral is subject to a KMS condition and unitarity constraints. To describe these, one typically works with the following (advanced/retarded) basis of operators:

$$\begin{aligned}\mathcal{O}_{\rm adv} &\equiv \mathcal{O}_R - \mathcal{O}_L\,,\\ \mathcal{O}_{\rm ret} &\equiv \left(\frac{\mathcal{O}_R+\mathcal{O}_L}{2}\right) - \frac{1}{2}\coth\left(\frac{i\beta\partial_t}{2}\right)(\mathcal{O}_R-\mathcal{O}_L)\,.\end{aligned} \tag{3.4.4}$$

This choice makes manifest the following constraints:

- The largest-time equations discussed earlier continue to hold in the form

$$\left\langle \mathcal{T}_{\mathcal{C}}\ \mathcal{O}_{\rm adv}(t_f)\,\mathcal{O}_{I_1}(t_1)\cdots\mathcal{O}_{I_n}(t_n)\right\rangle_\beta = 0 \qquad (t_f > t_1,\ldots t_n)\,. \tag{3.4.5}$$

- In addition, by virtue of the conjugation property (3.4.2) one has another constraint, the thermal smallest time equation, or equivalently the KMS condition

$$\left\langle \mathcal{T}_{\mathcal{C}}\ \mathcal{O}_{\rm ret}(t_i)\,\mathcal{O}_{I_1}(t_1)\cdots\mathcal{O}_{I_n}(t_n)\right\rangle_\beta = 0 \qquad (t_i < t_1,\ldots t_n)\,. \tag{3.4.6}$$

This follows from $\rho_\beta\,\mathcal{O}(t_i) = \mathcal{O}(t_i - i\beta)\rho_\beta = {\rm e}^{-i\beta\partial_{t_i}}\mathcal{O}(t_i)\rho_\beta$.

For illustration, consider two-point functions. The matrix of all possible SK two-point functions in this basis is

$$\begin{aligned}\mathbb{G}_\beta(t,t') &\equiv \begin{pmatrix}\langle\mathcal{T}_{\mathcal{C}}\,\mathcal{O}_{\rm ret}(t)\mathcal{O}_{\rm ret}(t')\rangle & \langle\mathcal{T}_{\mathcal{C}}\,\mathcal{O}_{\rm ret}(t)\mathcal{O}_{\rm adv}(t')\rangle\\ \langle\mathcal{T}_{\mathcal{C}}\,\mathcal{O}_{\rm adv}(t)\mathcal{O}_{\rm ret}(t')\rangle & \langle\mathcal{T}_{\mathcal{C}}\,\mathcal{O}_{\rm adv}(t)\mathcal{O}_{\rm adv}(t')\rangle\end{pmatrix}\\ &= \begin{pmatrix}0 & iG^R(t-t')\\ iG^A(t-t') & 0\end{pmatrix}\,,\end{aligned} \tag{3.4.7}$$

where

$$\begin{aligned}iG^R(t) &= \Theta(t)\langle[\mathcal{O}(t),\mathcal{O}(0)]\rangle_\beta\\ iG^A(t) &= -\Theta(-t)\langle[\mathcal{O}(t),\mathcal{O}(0)]\rangle_\beta\,.\end{aligned} \tag{3.4.8}$$

We should note that the KMS sliding rule directly leads to the fluctuation-dissipation relation, for

$$\begin{aligned} &\mathrm{Tr}(\rho_\beta\, B(0)\, A(t)) = \mathrm{Tr}(\rho_\beta\, A(t - i\,\beta) B(0)) \\ \implies\quad &\langle\{A(\omega_1), B(\omega_2)\}\rangle_\beta = \coth(\frac{\beta\,\omega_1}{2})\ \langle[A(\omega_1), B(\omega_2)]\rangle_\beta\ . \end{aligned} \tag{3.4.9}$$

At higher points the fluctuation dissipation relations will mix the correlators from across OTO contours differing by one timefold. Generically, all the $n!$ correlators can be broken up into KMS orbits of length n. The number of independent thermal n-point function is therefore $(n-1)!$. It is rather non-trivial combinatorial problem to decompose $(n-1)!$ into equivalence classes under both unitarity and KMS rules, analogous to (3.1.25). This was done in [10].

To give just one example, we noticed that the set of three-point Wightman functions involved both 1-OTO and 2-OTO observables. However, all the 2-OTO correlators can be related in this case to 1-OTO correlators using the KMS condition. The first independent set of correlators whose KMS orbits are confined to 2-OTO contours occur for four-point functions. The chaos correlator, which we discuss below, is the paradigmatic example of this.

3.4.2 One-Sided Thermal Basis

Define the one-sided thermal k-OTO path integral as

$$\mathcal{Z}^\beta_{\text{one-sided}}[J_1, \ldots, J_{2k}] = \mathrm{Tr}\left[U^\dagger[J_{2k}]U[J_{2k-1}]\cdots U^\dagger[J_2]U[J_1]\,\rho_\beta\right]. \tag{3.4.10}$$

This path integral is subject to a single KMS condition and $2k-1$ unitarity conditions. To describe these, we work with the following generalized retarded/advanced basis:

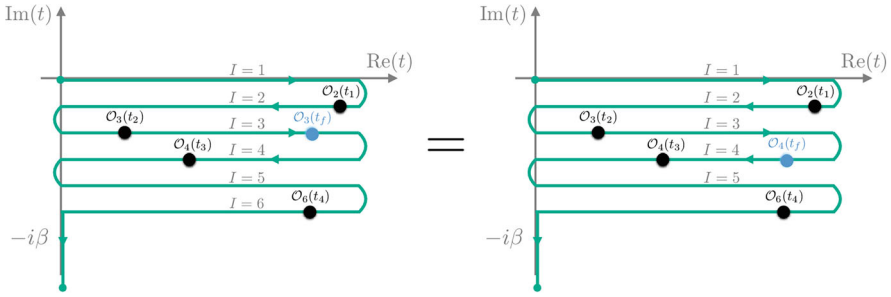
$$\begin{aligned} \mathcal{O}_{\text{adv}(\ell)} &\equiv \mathcal{O}_{\ell+1} - \mathcal{O}_\ell \qquad (\ell = 1, \ldots, 2k-1)\,, \\ \mathcal{O}_{\text{ret}} &\equiv \left(\frac{\mathcal{O}_{2k} + \mathcal{O}_1}{2}\right) - \frac{1}{2}\coth\left(\frac{i\beta\partial_t}{2}\right)(\mathcal{O}_{2k} - \mathcal{O}_1)\ , \end{aligned} \tag{3.4.11}$$

We refer to $\mathcal{O}_{\text{adv}(\ell)}$ as the set of $2k-1$ *advanced* combinations and $\mathcal{O}_r$ as the *retarded* combination. This basis again diagonalizes the unitarity and KMS condition for the one-sided path integral; that is, we have the following constraints:

- There are k largest-time equations of the form

$$\left\langle \mathcal{T}_\mathcal{C}\ \mathcal{O}_{\text{adv}(\ell=2m-1)}(t_f)\,\mathcal{O}_{I_1}(t_1)\cdots\mathcal{O}_{I_n}(t_n)\right\rangle_\beta = 0 \qquad (t_f > t_1, \ldots t_n)\,, \tag{3.4.12}$$

where $m = 1, \ldots, k$ and $\mathcal{O}_{I_i}$ are arbitrary operator combinations on the contours, and $\mathcal{T}_\mathcal{C}$ denotes k-OTO contour ordering. These constraints follow from unitarity: $U^\dagger[J_{\ell+1}]U[J_\ell] = 1$ when $J_{\ell+1} = J_\ell$ since $\mathcal{O}^{(\ell)}_{\text{adv}}$ is sourced by $\frac{1}{2}(J_{\ell+1} + J_\ell)$. Note that, strictly speaking, $\mathcal{O}_{\text{adv}(\ell)}(t_f)$ does not need to be to the future of *all* other operators, but only to the future of those operators that have a leg on contour segments ℓ or $\ell+1$. In other words, the largest time equation can be read as a 'sliding rule', which moves operators around a *future turning point* of the contour. For example:



- There are $k-1$ smallest-time equations,

$$\left\langle \mathcal{T}_\mathcal{C}\, \mathcal{O}_{\text{adv}(\ell=2m)}(t_i)\, \mathcal{O}_{I_1}(t_1) \cdots \mathcal{O}_{I_n}(t_n) \right\rangle_\beta = 0 \qquad (t_i < t_1, \ldots t_n)\,, \tag{3.4.13}$$

  where $m = 1, \ldots, k-1$. These follow again from unitarity as above. Again, t_i only needs to be the smallest time among those insertions that have legs on contour segments ℓ or $\ell+1$. Pictorially, this is a 'sliding rule' that takes operators around a *past turning point* of the contour.
- There is one KMS condition,

$$\left\langle \mathcal{T}_\mathcal{C}\, \mathcal{O}_{\text{ret}}(t_i)\, \mathcal{O}_{I_1}(t_1) \cdots \mathcal{O}_{I_n}(t_n) \right\rangle_\beta = 0 \qquad (t_i < t_1, \ldots t_n)\,. \tag{3.4.14}$$

  This follows the same way as in the 1-OTO case. Inside a k-OTO path ordered expectation value, we can also write the relevant condition as: $\mathcal{O}_{2k}(t_i) = \mathcal{O}_1(t_i - i\beta)$, assuming t_i is smaller than any other time of operator insertions with legs on contour segments 1 or $2k$.

3.4.3 Spectral Representation

There is a natural basis for thermal correlators that takes into account the KMS constraints and expresses the correlators in each KMS orbit in terms of a spectral function. From our earlier discussion, this should imply that there is a single spectral function for two- and three-point functions, and a pair of spectral functions for four-

point functions. The construction is achieved by extracting the statistical factors (Bose-Einstein/Fermi-Dirac) from the correlators. This construction is explained in [11] for generic k-OTOs. We will compile some essential features that we will use in our discussion later.

Since thermal two-point functions are fixed by a single spectral function, introduce

$$\int \frac{\mathrm{d}^{\mathrm{D}} p}{(2\pi)^{\mathrm{D}}}\, \varrho(p)\, \mathrm{e}^{i\, p\cdot(x_1-x_2)} = \langle [\mathcal{O}(x_1), \mathcal{O}(x_2)] \rangle_\beta \ . \tag{3.4.15}$$

Using this we can represent the contour-ordered two-point functions as

$$\langle \mathcal{T}_{\mathcal{C}}\, \mathcal{O}(x_1)\, \mathcal{O}(x_2) \rangle = \Theta_{12}\, M(x_1, x_2) + \text{permutations}\,, \tag{3.4.16}$$

where $M(x_1, x_2)$ is the matrix of Wightman correlators. It can be reconstructed from the spectral function as

$$M(x_1, x_2) = \int \frac{\mathrm{d}^{\mathrm{D}} p_1 \mathrm{d}^{\mathrm{D}} p_2}{(2\pi)^{2\mathrm{D}}}\, \varrho(p_1 - p_2) \begin{pmatrix} -1 \\ -1 \end{pmatrix} \mathrm{e}^{i\, p_1\cdot x_1} \otimes \begin{pmatrix} n_B(\omega_2) \\ 1 + n_B(\omega_2) \end{pmatrix} \mathrm{e}^{i\, p_2\cdot x_2}\,, \tag{3.4.17}$$

where n_B is the Bose-Einstein statistical factor

$$n_B(\omega) = \frac{1}{\mathrm{e}^{\beta\omega} - 1}\,. \tag{3.4.18}$$

Similar expressions can be obtained at higher points. For three-point functions, the spectral function is

$$\int \left[\prod_{i=1}^{3} \frac{\mathrm{d}^{\mathrm{D}} p_i}{(2\pi)^{\mathrm{D}}}\right] \varrho(p_1, p_2, p_3)\, \mathrm{e}^{i\, \sum_{i=1}^{3} p_i\cdot x_i} = \langle [[\mathcal{O}(x_1), \mathcal{O}(x_2)]\,, \mathcal{O}(x_3)] \rangle_\beta \ . \tag{3.4.19}$$

For four-point functions we have

$$\begin{aligned} &\int \left[\prod_{i=1}^{4} \mathrm{e}^{i\, p_i\cdot x_i} \frac{\mathrm{d}^{\mathrm{D}} p_i}{(2\pi)^{\mathrm{D}}}\right] \varrho_1(p_1, p_2, p_3, p_4) \\ &\quad = \langle [[[\mathcal{O}(x_1), \mathcal{O}(x_2)]\,, \mathcal{O}(x_3)]\,, \mathcal{O}(x_4)] \rangle_\beta \ , \\ &\int \left[\prod_{i=1}^{4} \mathrm{e}^{i\, p_i\cdot x_i} \frac{\mathrm{d}^{\mathrm{D}} p_i}{(2\pi)^{\mathrm{D}}}\right] \varrho_2(p_1, p_2, p_3, p_4) \\ &\quad = \langle [\mathcal{O}(x_1), \mathcal{O}(x_2)]\ [\mathcal{O}(x_3), \mathcal{O}(x_4)] \rangle_\beta \ . \end{aligned} \tag{3.4.20}$$

Finally, we introduce another basis closely related to the advanced/retarded basis mentioned earlier, but with the salubrious feature of diagonalizing the KMS

constraints. We define the future and past combinations of sources to be

$$J_F = -((1+n_B)\,J_R - n_B\,J_L)\,, \qquad J_P = -n_B\,(J_R - J_L)\,. \tag{3.4.21}$$

The operators can likewise be rotated in a similar fashion.

3.5 Schwinger–Keldysh Contours in Gravity: grSK

Mukund Rangamani

We now turn to the question of how to implement the Schwinger–Keldysh path integral contour in holographic settings, as indicated in our discussion of the open QFT paradigm. For this discussion, we will focus primarily on the thermal state of the holographic theory. Furthermore, we will focus exclusively on the Schwinger-Keldysh contour, for which there is now a well understood gravitational prescription.

We first recall that the thermal state can be prepared by a Euclidean path integral, where the CFT is placed on a geometry with a compact time coordinate. As in our previous discussion, one opens up the contour around $t_E = 0$ to allow for the real-time timefolded evolution.

The Schwinger–Keldysh contour of the field theory maps in the gravitational setting to a boundary condition for the semiclassical quantum gravity path integral. The natural question involves identifying the corresponding bulk geometry, which we recall has to be a saddle point of the gravitational path integral. We will follow the prescription of [12], as explained in [8]. We should note important previous work on the subject [13, 14] and in particular [15].

The thermal state preparation is achieved by a non-trivial saddle of the gravitational path integral. The geometry is the Euclidean black hole spacetime, with the time circle fixed (in an appropriate conformal frame) at the AdS boundary. This circle shrinks to zero in the interior of the spacetime, with the locus of zero proper size corresponding to the bifurcation surface of the event horizon in the Lorentzian continuation. The spacetime has the topology of a disc (often referred to as the cigar geometry). We open up this geometry in the vicinity of $t_E = 0$ and patch in two copies of the Lorentzian spacetime, which are glued together along their respective future horizons, see Fig. 3.4.

It is useful to write an explicit set of coordinates and present this solution as a complex two-sheeted geometry. To do so, consider the metric in ingoing Eddington-Finkelstein coordinates

$$\mathrm{d}s^2 = -r^2\,f(r)\,\mathrm{d}v^2 + 2\,\mathrm{d}v\,\mathrm{d}r + r^2\,\mathrm{d}\mathbf{x}^2\,, \qquad f(r) = 1 - \frac{r_+^{\mathrm{D}}}{r^{\mathrm{D}}}\,. \tag{3.5.1}$$

The coordinate v is identified with the time coordinate t on the boundary of the spacetime $r \to \infty$, and follows the Schwinger–Keldysh contour. The radial

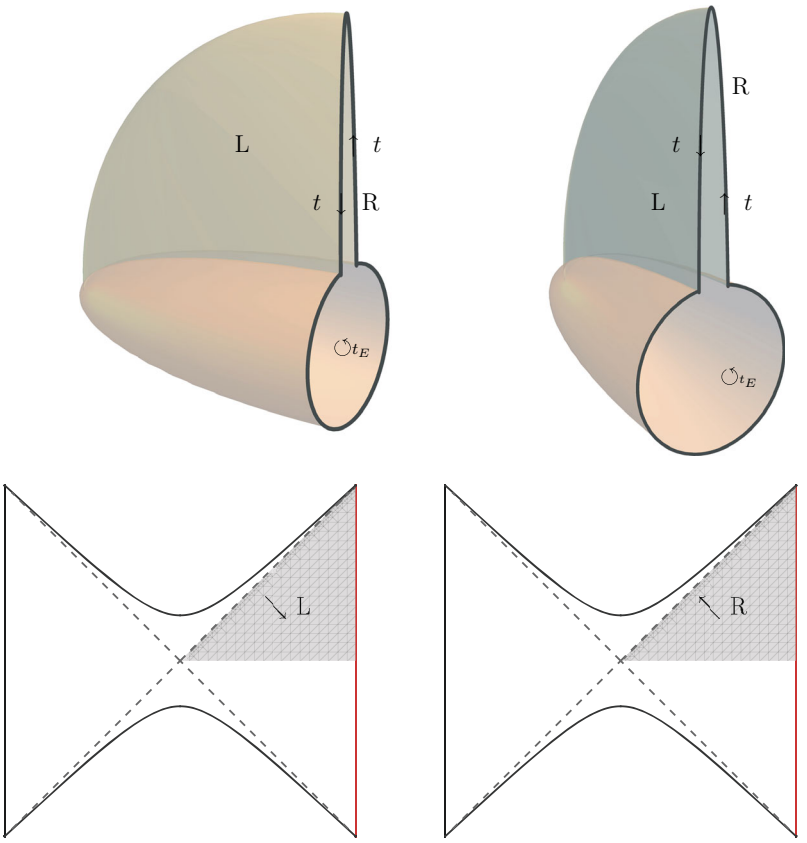


Fig. 3.4 The two-sheeted complex grSK geometry shown from two different perspectives. On the top left we display the boundary thermal SK contour, which is filled in the Euclidean portion by the Euclidean black hole geometry (the cigar) and in the Lorentzian section by two copies of the domain of outer communication of the Lorentzian black hole spacetime. The top right panel displays the bulk perspective to emphasize the smooth join of the two sheets of the Lorentzian section. On the bottom panel, we illustrate the Lorentzian sections of the geometry on the Schwarzschild-AdS$_{D+1}$ Penrose diagram. with the regions pertaining to the L and R sheets of the grSK spacetime shaded. Reprinted under CC-BY-4.0 license from [8]. © 2020, The Author(s)

coordinate in the bulk will be complexified, and we will give a contour prescription to choose a codimension-1 slice.

Operationally, we, in fact, upgrade the radial tortoise coordinate to a complex variable, which we refer to as the *mock tortoise* coordinate, ζ. We define this coordinate by the differential relation:

$$\frac{\mathrm{d}r}{\mathrm{d}\zeta} = \frac{i\,\beta}{2}\, r^2\, f(r)\,, \tag{3.5.2}$$

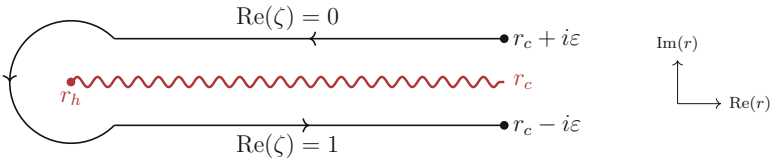


Fig. 3.5 The complex r plane with the locations of the two boundaries and the horizon marked. The grSK contour is a codimension-1 surface in this plane (drawn at fixed v). The direction of the contour is indicated counterclockwise, circling the branch point at the horizon

where $\beta = \frac{4\pi}{\mathrm{D} r_+}$ is the inverse temperature of the black hole. While we have indicated the explicit expressions for the case of a Schwarzschild-AdS$_{\mathrm{D}+1}$ black hole, the construction straightforwardly generalizes to any non-extremal black hole, see [9].

By convention, we choose one of the sheets to have a vanishing real part and the other to have a unit real part. This fixes the normalization used to define ζ.

$$\lim_{r\to\infty} \zeta(r + i\,\varepsilon) = 0\,, \qquad \lim_{r\to\infty} \zeta(r - i\,\varepsilon) = 1\,. \tag{3.5.3}$$

A section of geometry in the mock tortoise complex plane is illustrated in Fig. 3.5. The complexified metric defining the grSK geometry is

$$\mathrm{d}s^2 = -r^2\, f(r)\, \mathrm{d}v^2 + i\,\beta\, r^2\, f(r)\, \mathrm{d}v\, \mathrm{d}\zeta + r^2\, \mathrm{d}\mathbf{x}^2\,, \qquad f(r) = 1 - \frac{r_+^{\mathrm{D}}}{r^{\mathrm{D}}}\,. \tag{3.5.4}$$

One can now give a simple algorithm for computing real-time observables in the holographic CFT using Witten diagrams on the grSK geometry.

- To begin with one places the gravitational theory on the grSK contour, writing for the bulk action a contour integral

$$S_{\mathrm{bulk}} = \oint \mathrm{d}\zeta \int \mathrm{d}^{\mathrm{D}}x\, \sqrt{-g}\, \mathcal{L}[g_{AB}, \Phi]\,. \tag{3.5.5}$$

  Here, Φ denotes the set of bulk fields associated to the CFT operator $\mathcal{O}$, and g_{AB} the bulk metric is dual to the CFT stress tensor.
- The main task is to determine the propagators for the linearized fluctuations of the fields Φ (and metric) around the grSK saddle. To this end, consider the linear wave equation in the original black hole spacetime in ingoing coordinates (3.5.1). Working in Fourier domain with frequencies and momenta denoted ω and $\mathbf{k}$, one solves the wave equation. The boundary conditions to be used are ingoing boundary conditions at the horizon (which ensure regularity) and requiring the field to asymptote to an appropriate source at the conformal boundary. We will

refer to this as the ingoing boundary to bulk propagator $G_{\rm in}$

$$\lim_{r\to\infty} G_{\rm in} = 1\,, \qquad \left.\frac{\mathrm{d}G_{\rm in}}{\mathrm{d}r}\right|_{r_+} = 0\,. \tag{3.5.6}$$

- The ingoing boundary to bulk propagator, being regular at the horizon, does not see the branch cut inherent in the definition of ζ. Therefore, one can immediately write down the ingoing grSK propagator $G_{\rm in}(\omega, \mathbf{k}, \zeta)$ with suitable replacements.
- The second propagator that we need is the outgoing propagator. To this end we exploit the time-reversal involution of the grSK geometry and deduces that it is given by frequency reversing the ingoing propagator and multiplying with a radial Boltzmann factor

$$G_{\rm out}(\omega, \mathbf{k}, \zeta) \equiv G_{\rm in}(-\omega, \mathbf{k}, \zeta)\, \mathrm{e}^{-\beta\,\omega\,\zeta}\,. \tag{3.5.7}$$

- The solution to the field Φ with L and R Schwinger–Keldysh sources on the respective boundaries then can be written solely in terms of the ingoing propagator. We have

$$\begin{aligned}\Phi(\zeta, \omega, \mathbf{k}) &= G_{\rm in}(\omega, \mathbf{k}, \zeta)\left((1+n_\omega)\,J_R - n_\omega\,J_L\right)\\ &\quad - G_{\rm in}(-\omega, \mathbf{k}, \zeta)\,\mathrm{e}^{\beta\,\omega\,(1-\zeta)}\,n_\omega\Big(J_R - J_L\Big)\,,\\ &= -G_{\rm in}(\zeta, \omega, \mathbf{k})\,J_F + G_{\rm in}(-\omega, \mathbf{k}, \zeta)\,\mathrm{e}^{\beta\omega\,(1-\zeta)}\,J_P\,.\end{aligned} \tag{3.5.8}$$

- The final ingredient is the bulk-bulk propagator, which can be obtained using the variation of parameters trick. One takes advantage of the solutions of the homogeneous wave equation to construct the solutions to the inhomogeneous one. This can be efficiently expressed in terms of a linear combination of boundary-bulk Green's functions that are normalizable at one or the other boundary of the grSK geometry. Denoting these as $G_R(\zeta, \omega, \mathbf{k})$ and $G_L(\zeta, k)$, respectively, we have (setting $k = (\omega, \mathbf{k})$ for simplicity)

$$\begin{aligned}G_R(\zeta, k) &= \mathrm{e}^{\beta\,\omega}\,n_B(\omega)\,(G_{\rm in}(\zeta, k) - G_{\rm out}(\zeta, k))\,,\\ G_L(\zeta, k) &= -n_B(\omega)\left(G_{\rm in}(\zeta, k) - \mathrm{e}^{\beta\omega}\,G_{\rm out}(\zeta, k)\right).\end{aligned} \tag{3.5.9}$$

The function G_R has a source on the right boundary ($\zeta = 1$), while G_L has a source on the left boundary ($\zeta = 0$), and they are respectively normalizable at the other end viz.,

$$\lim_{\zeta\to 0}\{G_R, G_L\} = \{0, 1\}\,, \qquad \lim_{\zeta\to 1}\{G_R, G_L\} = \{1, 0\}\,. \tag{3.5.10}$$

The bulk-bulk propagator can be seen to be given by

$$G_{\text{bulk}}(\zeta,\zeta';k) = \mathcal{N}(k)\, e^{\beta\omega\zeta'}\Big[\Theta(\zeta-\zeta')\, G_L(\zeta,k)\, G_R(\zeta',k) + \Theta(\zeta'-\zeta)\, G_L(\zeta',k)\, G_R(\zeta,k)\Big]. \tag{3.5.11}$$

Here $\Theta(\zeta-\zeta')$ is a contour-ordered step function along the contour depicted in Fig. 3.5. The prefactor can be obtained from the Wronskian of the two linearly independent solutions to the homogeneous equation, which we have taken to be the left- and right-normalizable boundary-bulk propagators.

Armed with these propagators, one simply writes down Witten diagrams for the boundary correlators with the prescribed sources. Each diagram will have a set of ingoing and outgoing propagators and bulk-to-bulk propagators sewn together at the vertices. This defines the integrand of the Witten diagram, and the integration is over the grSK contour sketched in Fig. 3.5. As demonstrated in [8, 9, 16] each such diagram can be expressed as an integral over a single copy of the black hole exterior, along the real contour running from r_+ to ∞ of an integrand, which is a sum of (multiple) discontinuities of the contour integrand.[4]

3.6 Applications of k-OTO Path Integrals

Felix Haehl

We will now discuss some concrete applications of k-OTO path integrals, focusing in particular on $k \geqslant 2$. The goal is to understand some related physical phenomena, which are often of an information-theoretic nature.

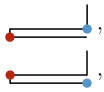
3.6.1 OTOCs in Thermal States

We begin with a review of out-of-time-order correlation functions and their relation to quantum chaos. The basic ideas were first discussed in [17–20].

3.6.1.1 Quantum Butterfly Effect

We would like to consider the two states

$$|\Psi\rangle \equiv V(-T,0)W(0,x)|\beta\rangle \equiv e^{-iHT}V(0,0)e^{iHT}W(0,x)|\beta\rangle = \quad ,$$
$$|\Phi\rangle \equiv W(0,x)V(-T,0)|\beta\rangle \equiv W(0,x)e^{-iHT}V(0,0)e^{iHT}|\beta\rangle = \quad , \tag{3.6.1}$$



[4] This is assuming that the vertex functions are simple – in certain cases there are also localized contributions at the horizons which originate from derivative interactions. See the aforementioned references for more details.

where $|\beta\rangle = \mathrm{e}^{-\beta H/2}|0\rangle$ is the square root of the thermal density matrix such that $\langle\beta|(\cdots)|\beta\rangle \equiv \mathrm{Tr}[(\cdots)\rho_\beta]$. We expect that the two states are generically very similar when the time separation $T \ll \beta$. However, if H is a suitably *chaotic* Hamiltonian, then the overlap becomes small for large time separations. By the overlap, we mean the following 2-OTO correlation function:

$$\begin{aligned}\langle\Phi|\Psi\rangle &= \mathrm{Tr}\left[V^\dagger(-T,0)W^\dagger(0,x)V(-T,0)W(0,x)\,\rho_\beta\right] = \\ &\sim \langle W^\dagger W\rangle_\beta \langle V^\dagger V\rangle_\beta \underbrace{\left[1-\frac{c_0}{N}\,\mathrm{e}^{\kappa(v)T}+\mathcal{O}(N^{-2})\right]}_{\to 0\ (T\to\infty)}\end{aligned} \tag{3.6.2}$$

where κ is the Lyapunov exponent. In general, the Lyapunov exponent depends on the spatial separation between V and W. Specifically, it is expected to be a function of the 'velocity' $v = x/T$, or equivalently of the relative spatial momentum. We will suppress this spatial dependence for now and get back to it later.

Let us now be more precise about the meaning of the above. We canonically purify the system by doubling the Hilbert space $\mathcal{H} \to \mathcal{H}_L \otimes \mathcal{H}_R$. We use a convention where the time in the R-Hilbert space is aligned with global time, $t_R = t$, and the time in the L-Hilbert space runs in the opposite direction, $t_L = -t$. The purification of the thermal density matrix corresponds to a state $|\mathrm{TFD}\rangle \in \mathcal{H}_L \otimes \mathcal{H}_R$. Let us discuss three closely related correlation functions in the doubled system (see, for example, [21]), which make different features of our discussion manifest (square brackets).

1. One-Sided Correlator [OTO Path Integral] The states we compare can be chosen in various ways; a conveninent choice is to represent all operators in the R-Hilbert space:

$$|\Psi_1\rangle \equiv V_R(-T)W_R(0)|\mathrm{TFD}\rangle\,, \qquad |\Psi_2\rangle \equiv W_R(0)V_R(-T)|\mathrm{TFD}\rangle\,, \tag{3.6.3}$$

such that their overlap reads as

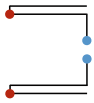
$$\langle\Psi_2|\Psi_1\rangle = \langle\mathrm{TFD}|V_R^\dagger W_R^\dagger V_R W_R\,|\mathrm{TFD}\rangle = \mathrm{Tr}\left[V^\dagger W^\dagger V W \rho_\beta\right]. \tag{3.6.4}$$

Written in this way, the overlap $\langle\Psi_2|\Psi_1\rangle$ is a thermal 4-point 2-OTO correlation function of the type discussed in previous sections. It compares the effect of acting with $W(0)$ first followed by $V(T)$, versus acting in the opposite order. In a large N system and for 'simple' operators V, W, the difference between these two states is initially negligible, but can become large as T grows. This is the basic feature of the quantum butterfly effect.

2. Overlap of 'in' and 'out' States [Scattering Experiment] The one-sided OTOC is closely related to the overlap of an 'in' state on the time slice at an early

global time and an 'out' state at late time:[5]

$$|\Psi_{\text{in}}\rangle \equiv W_L(0)V_R(-T)|\text{TFD}\rangle = \quad |\Psi_{\text{out}}\rangle \equiv V_L(T)W_R(0)|\text{TFD}\rangle = \tag{3.6.5}$$



such that

$$\langle\Psi_{\text{out}}|\Psi_{\text{in}}\rangle = \langle\text{TFD}|V_L^\dagger W_L W_R^\dagger V_R|\text{TFD}\rangle = \text{Tr}\left[WV^\dagger\rho_\beta^{1/2}W^\dagger V\rho_\beta^{1/2}\right] = \tag{3.6.6}$$

The labels 'in' and 'out' refer to whether the states are created in the past or future with respect to global time. This version of the OTOC is most easily interpreted as a scattering problem. The 'in' and 'out' states are defined on time slices, i.e., they each involve two spacelike separated operator insertions. We can then describe their difference in terms of a 2-to-2 scattering phase; see, e.g., [22].

3. Two-Point Function in a Perturbed State [Probing a Shockwave State] Finally, consider the state

$$|\Phi_V\rangle \equiv V_R(0)|\text{TFD}\rangle\,, \tag{3.6.7}$$

and write the left-right two-point function of W:

$$\begin{aligned}\langle\Phi_V|W_L^\dagger(-T)W_R(T)|\Phi_V\rangle &= \langle\text{TFD}|W_L^\dagger V_R^\dagger W_R V_R|\text{TFD}\rangle \\ &= \text{Tr}\left[W^\dagger\rho_\beta^{1/2}V^\dagger W V\rho_\beta^{1/2}\right].\end{aligned} \tag{3.6.8}$$

This analytic continuation of the OTOC cannot be written as the overlap of an 'in' and 'out' state. It should be thought of as a two-point probe correlator evaluated in the state that corresponds to a thermal state perturbation by V.

It is instructive to compute OTOCs using holography. Doing so properly should involve a generalization of the setup described in Sect. 3.5 to geometries with boundary conditions appropriate for the OTO path integral ('grOTO geometry'). This has not yet been studied for the correlators described here. However, the calculation is easy in a certain approximation. For instance, the state (3.6.7) evolved to a sufficiently late time slice can be approximated by an eternal black hole geometry perturbed by a null shockwave that travels from the right boundary

[5] Recall $t_L = -t$, which means $V_L(t_L = T)$ corresponds to an insertion at global time $-T$. Contour pictures are drawn in the complex 'global' t-plane.

towards the black hole. That is, the gravitational backreaction (graviton exchanges on the grOTO geometry) is effectively reorganized into a gravitational shockwave. The OTOC (3.6.8) is then computed by a late-time left-right correlator in an eternal black hole geometry perturbed by a shockwave [17].

3.6.1.2 Six-Point OTOCs and Shockwave Collisions

Higher-point OTOCs play a natural role when one tries to measure detailed properties of states that are perturbed with respect to some reference state.

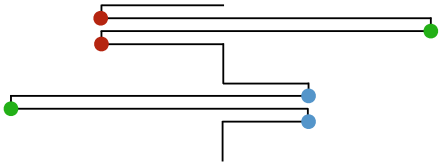
A 3-OTO application can be motivated as follows [23–25]. Consider the perturbed thermofield double state:

$$|\Phi_{VW}\rangle \equiv W_L(-t_{wL})V_R(-t_{vR})|\text{TFD}\rangle\,. \tag{3.6.9}$$

This state has an interesting interpretation in holography: for large t_{vR} and t_{wL}, it corresponds approximately to the eternal AdS black hole perturbed by two shockwaves traveling from the left and right boundaries, respectively, in almost null directions (see figures below). The two shocks may scatter behind the black hole horizon. We can try to characterize this scattering process and how it affects the post-collision geometry. A natural probe would be a left-right two-point function on a late time slice $t_0 \gg t_{wL}, t_{vR}$:

$$\begin{aligned}&\langle\Phi_{VW}|\mathcal{O}_L(t_0)\mathcal{O}_R(t_0)|\Phi_{VW}\rangle\\ &\quad= \text{Tr}\left[W^\dagger(t_{wL})\mathcal{O}(-t_0)W(t_{wL})\,\rho_\beta^{1/2}\,V^\dagger(-t_{vR})\mathcal{O}(t_0)V(-t_{vR})\,\rho_\beta^{1/2}\right],\end{aligned} \tag{3.6.10}$$

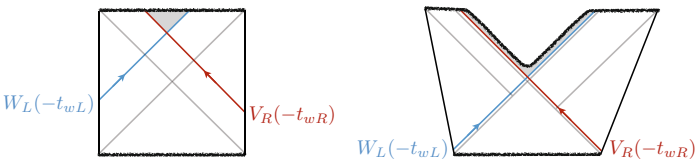
where we translated the two-sided correlator into a trace (3-OTO path integral) over a single Hilbert space. On the contour, it looks schematically as follows:



Due to the isometries of the eternal black hole geometry, the result of this calculation should depend on the total separation $\Delta t = t_{wL} + t_{vR}$. One finds that for small Δt the six-point function cluster-decomposes into two-point functions to a good approximation. However, when *both* V and W each had enough time to scramble, the correlator will deviate significantly from the factorized result. One finds [24]:

$$\langle\Phi_{VW}|\mathcal{O}_L(t_0)\mathcal{O}_R(t_0)|\Phi_{VW}\rangle \approx \langle V^\dagger V\rangle_\beta\langle W^\dagger W\rangle_\beta\langle\mathcal{O}\mathcal{O}\rangle_\beta\left[\frac{1}{1+\alpha\, G^2\Delta_V\Delta_W\, e^{\Delta t}}\right]^{2\Delta_\mathcal{O}} \tag{3.6.11}$$

with some theory- and dimension-dependent number α and the Newton constant G. Importantly, this result deviates from the product of two-point functions by a large amount after *two* scrambling times, i.e., when $\Delta t \sim t_{*W} + t_{*V}$ with $t_{*\mathcal{O}} \equiv \log(\sqrt{\alpha} G \Delta_{\mathcal{O}})$.[6] This deviation indicates that the gravitational interaction between the two shockwaves is very disruptive. Indeed, the decorrelation of the left-right two-point function can be interpreted as the disappearance of spacetime connecting the two sides. A more detailed gravitational analysis indicates that the total spacetime volume within the causal future of the collision point rapidly decreases to an exponentially small value because of the strong non-linear gravitational backreaction. The two situations can be pictured as follows:[7]



$(\Delta t < t_{*W}, t_{*V})$ $(\Delta t > t_{*W} + t_{*V})$

Can we understand this process using a more conventional language appropriate to describe the 2-to-2 scattering process? An immediate obstacle is the ill-definedness of asymptotic out states due to the spacetime singularity. Initially (for small Δt) this might not be a serious issue because the singularity is far from the collision point; but for large Δt, the singularity "bends down" exponentially close to the collision point. Characterizing the collision in terms of an S-matrix now seems quite problematic. It is no coincidence that the six-point OTOC considered above is not naturally written as the overlap of an in- and an out-state. Relatedly, the four-point OTOC (which *can* be written as such an overlap, see (3.6.6)) is not sensitive to the change of behavior at two scrambling times. It is an interesting question to investigate how to better understand this process in the language of the S-matrix.

3.6.2 The Effective Theory of Chaos and Eikonal Scattering

In chaotic systems, it seems to be the case that the SK (or OTO) path integral computing out-of-time order correlators is dominated by few collective modes, which we will refer to as "scramblons". We will now define these modes and

[6] For $t_0 \ll t_{vR}, t_{wL}$ the six-point OTOC provides a different refined measure of scrambling, see [26, 27].

[7] We are glossing over some details here: initially, as we increase t_{wR} and t_{wL} from zero, the gray spacetime region grows (roughly linearly). After one scrambling time, the gravitational backreaction becomes strong and the singularity "bends down" in such that the grey region begins to shrink. After two scrambling times, it has shrunk to an exponentially small value, as shown on the right [25].

their effective description, illustrating how their dynamics gives rise to the butterfly effect. It is interesting to note that scramblons live inherently on the real-time contour [28–32].

3.6.2.1 Generalities

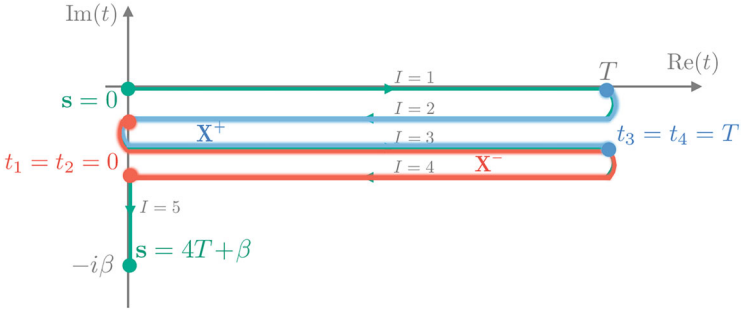
For concreteness, let us consider the following OTOC:

$$\mathrm{OTOC}(T,x) \equiv \frac{\mathrm{Tr}\left[V^\dagger(0,0)W^\dagger(T,x)V(0,0)W(T,x)\,\rho_\beta\right]}{\mathrm{Tr}\left[V^\dagger V\rho_\beta\right]\mathrm{Tr}\left[W^\dagger W\rho_\beta\right]}\,. \tag{3.6.12}$$

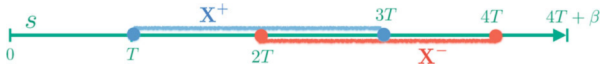
To compute this, we parameterize the thermal 2-OTO contour using a *contour time* coordinate $s \in [0, 4T+\beta]$ such that the usual complex time t takes the following values along the contour:

$$t_I(s) = \begin{bmatrix} s & (0 \leqslant s \leqslant T)\,, \\ 2T-s & (T \leqslant s \leqslant 2T)\,, \\ s-2T & (2T \leqslant s \leqslant 3T)\,, \\ 4T-s & (3T \leqslant s \leqslant 4T)\,, \\ -i(s-4T) & (4T \leqslant s \leqslant 4T+\beta)\,, \end{bmatrix} \tag{3.6.13}$$

where $I = 1, \ldots, 5$ labels the five segments of the contour ($I = 5$ is Euclidean):



Conceptually, it can be useful to "unwrap" this contour, which makes the flow of contour time more manifest:



Consider now a (schematic) path integral along this contour:

$$\mathcal{Z}[J_I] = \int [D\phi_I]\ \exp\left\{iS[\phi_1,J_1] - iS[\phi_2,J_2] + iS[\phi_3,J_3] - iS[\phi_4,J_4] - S_E[\phi_5,J_5]\right\}\,, \tag{3.6.14}$$

where, for example, $S[\phi_1, J_1] = \int_0^T \mathrm{d}s \int \mathrm{d}^{D-1}x\, (\mathcal{L}[\phi_1] + J_1\phi_1)$, and so on.

In the above path integrals, ϕ_I is a stand-in for the various fields we need to integrate over. Consider the saddle point solution $\phi_\star$ relevant for the thermal state. In the rest of this section we will illustrate the following picture:

Effective Description of Scrambling Assume that the theory has many degrees of freedom and exhibits fast scrambling. Consider the saddle point corresponding to the thermal state. Then the OTO path integral computing (3.6.12) is dominated by the exchange of soft mode fluctuations $\delta\phi_I$ defined on the thermal OTO contour, which grow or decay exponentially in time. Their quadratic action becomes parametrically small around the scrambling time $t_* \sim \log N$.

Specifically, the fluctuations described above are of the form

$$\delta_+\phi_I(t_I, p) = X^+(p)\, \mathrm{e}^{-\kappa(p)t_I}\,, \qquad \delta_-\phi_I(t_I, p) = X^-(p)\, \mathrm{e}^{\kappa(p)\,(t_I - T)}\,. \tag{3.6.15}$$

where κ is the quantum Lyapunov exponent and $X^\pm(p)$ parametrize the strength of the fluctuation as a function of spatial momentum. These fluctuations individually have a vanishing quadratic action, but their interaction leads to an effective 'eikonal' action that is proportional to X^+X^- and exponentially small at late times:

$$i\,S_{\mathrm{eik}}[X^+, X^-] = N \int \mathrm{d}p\; c(\kappa)\, \mathrm{e}^{\frac{i\pi\kappa}{2} - \kappa T}\; X^+(p)X^-(-p)\,. \tag{3.6.16}$$

The exponential suppression $\mathrm{e}^{-\kappa T}$ means that the large N saddle point approximation breaks down for the contribution of these modes to the OTO path integral. We will account for this by performing the path integral over the modes $X^\pm$ fully. The phase factor $\mathrm{e}^{\frac{i\pi\kappa}{2}}$ in the eikonal action is a universal feature when $\kappa \neq 1$ and a manifestation of the KMS condition [30, 31].

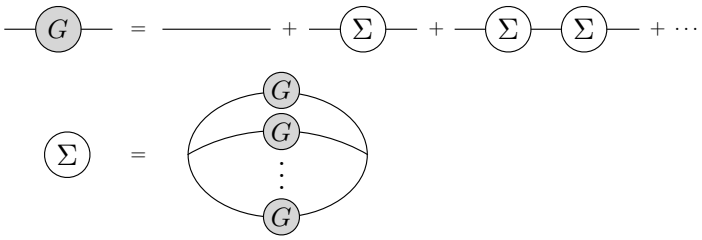
3.6.2.2 A Toy Model

Although we would ultimately like to understand the physics discussed here in theories such as $\mathcal{N} = 4$ SYM (let alone QCD), we will discuss a toy model instead, which exhibits some of the crucial features.

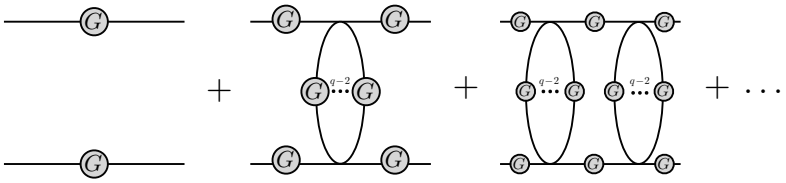
The Sachdev–Ye–Kitaev (SYK) model describes disordered interactions between N Majorana fermions [33, 34]. After disorder averaging, the large N model can be solved in terms of a collective variable $G(\tau_1, \tau_2)$ that is bilocal in time. It gives the mean-field value of the two-point functions averaged over all species of fermions and it satisfies the following Schwinger-Dyson equation:

$$\hat{G}(\omega_n) = -\frac{1}{i\omega_n + \hat{\Sigma}(\omega_n)}\,, \qquad \Sigma(\tau, \tau') = J^2\, G(\tau, \tau')^{q-1}\,, \tag{3.6.17}$$

where the self-energy $\Sigma(\tau,\tau') \equiv \sum_n e^{-i\omega_n(\tau-\tau')}\hat{\Sigma}(\omega_n)$ and $\omega_n = \frac{1}{2} + in$. Pictorially:



Since $G(\tau_1,\tau_2)$ is the full two-point function, we can use it to build higher-point functions. For example, a four-point turns out to take the following form:



At large J^2, the $i\omega_n$ can be neglected and the equations become invariant under time reparametrizations:

$$\begin{aligned} \tau &\mapsto f(\tau)\,, \\ G(\tau,\tau') &\mapsto [f'(\tau)f'(\tau')]^{\frac{1}{q}}\, G(f(\tau),f(\tau'))\,, \\ \Sigma(\tau,\tau') &\mapsto [f'(\tau)f'(\tau')]^{1-\frac{1}{q}}\, \Sigma(f(\tau),f(\tau'))\,. \end{aligned} \tag{3.6.18}$$

It is interesting to perform perturbation theory in $1/(\beta J)$ around the strict conformal limit. The $1/(\beta J)$ fluctuations lift the degeneracy and lead to a small action cost associated with any given reparametrization $f(\tau)$. The action is the simplest expression that preserves an $\mathrm{SL}(2,\mathbb{R})$ symmetry of the conformal saddle:

$$S = -C\int \mathrm{d}\tau\, \{f(\tau),\tau\}\,, \qquad C \propto \frac{N}{\beta J}\,. \tag{3.6.19}$$

This is vaguely similar to what happens in AdS/CFT: while the bulk theory at small G_N is approximated by classical gravity, there are some fluctuations around classical saddles that are enhanced, namely the massless stringy modes controlled by $\ell_s/\ell_{\mathrm{AdS}}$.

Scrambling in the Schwarzian Theory.
The Schwarzian theory (3.6.19) arises in the strong coupling limit of the SYK model, among other nearly conformal $(0+1)$-dimensional chaotic spin systems [34]. The Schwarzian path integral is over time reparametrizations $t \to f(t)$.

Let us now work on the thermal 2-OTO contour appropriate for computing the OTOC (following [31]). The saddle point corresponding to the thermal state is:

$$f_I(s) = \tanh\left(\frac{t_I(s)}{2}\right) \qquad (I = 1, \ldots, 5)\,. \tag{3.6.20}$$

The quadratic action of fluctuations $t_I(s) \to t_I(s) + \epsilon_I(t_I(s))$ is

$$i S_{\text{quad}} = \frac{iC}{2} \int_0^{4T+\beta} \mathrm{d}s\, \frac{\mathrm{d}t_I}{\mathrm{d}s}\, \epsilon_I(t_I(s)) \left(\partial_{t_I}^4 - \partial_{t_I}^2\right) \epsilon_I(t_I(s))\,. \tag{3.6.21}$$

From this action we can read off the real-time propagators of the Schwarzian mode along the contour. The contour-ordered propagators can all be expressed in terms of Schwarzian Green's functions [35]:

$$\begin{aligned} iG_\epsilon^R(t) &= \Theta(t)\,\langle[\epsilon(t),\,\epsilon(0)]\rangle_\beta = \frac{1}{iC}\,\Theta(t)\,(t - \sinh t) \sim -\frac{1}{2iC}\,\Theta(t)\,\mathrm{e}^t\,, \\ iG_\epsilon^A(t) &= -\,\Theta(-t)\,\langle[\epsilon(t),\,\epsilon(0)]\rangle_\beta = -\frac{1}{iC}\,\Theta(-t)\,(t - \sinh t) \sim -\frac{1}{2iC}\,\Theta(-t)\,\mathrm{e}^{-t}\,. \end{aligned} \tag{3.6.22}$$

The Schwarzian mode couples to external operators universally via time reparametrizations. This defines natural *bilocal* operators, which replace pairwise operator insertions:

$$\mathcal{O}_\Delta(t_I)\mathcal{O}_\Delta(t_J) \xrightarrow{f} \mathcal{B}_{IJ}^\Delta(t_I, t_J) = \left[-\frac{f_I'(t_I) f_J'(t_J)}{(f_I(t_I) - f_J(t_J))^2}\right]^\Delta, \tag{3.6.23}$$

which we can linearize as

$$\mathcal{B}_{IJ}^\Delta(t_I, t_J) = \langle\mathcal{O}_\Delta(t_I)\mathcal{O}_\Delta(t_J)\rangle_\beta \left[1 + \Delta\left[\epsilon_I'(t_I) + \epsilon_J'(t_J) - \frac{\epsilon_I(t_I) - \epsilon_J(t_J)}{\tanh\left(\frac{t_I - t_J}{2}\right)}\right] + \mathcal{O}(\epsilon^2)\right]. \tag{3.6.24}$$

In the OTOC, we consider these bilocal operators for the case where the two operators are inserted at the same time, only separated by a small imaginary

regulator that distinguishes the contours. In the above equations, this means $J = I + 2$ and $t_I = t$, $t_J = t - 2i\varepsilon$. The linearized bilocal becomes

$$\mathcal{B}^{\Delta}_{I,I+2}(t, t-2i\varepsilon) = \frac{\left[1+2\Delta\left(\partial_t \epsilon_{\text{av}(I,I+2)}(t) - \frac{\epsilon_{\text{dif}(I,I+2)}(t)}{2i\,\sin\varepsilon}\right)+\mathcal{O}(\epsilon^2)\right]}{(2\sin\varepsilon)^{2\Delta}}\,, \tag{3.6.25}$$

where we use a basis of the form

$$\epsilon_{\text{av}(I,J)}(t) = \frac{\epsilon_I(t)+\epsilon_J(t)}{2}\,, \qquad \epsilon_{\text{dif}(I,J)}(t) = \epsilon_I(t) - \epsilon_J(t)\,. \tag{3.6.26}$$

In the $\varepsilon \to 0$ regularization, the contribution due to the difference operator dominates. We can say that the external operator insertions act as sources for $\epsilon_{\text{dif}(I,I+2)}$. This means, for example, that the leading connected contribution to the OTOC is approximately

$$\text{OTOC}(T) = \langle VV\rangle_\beta \langle WW\rangle_\beta \left[1 - \frac{\Delta_V \Delta_W}{\sin^2\varepsilon}\,\left\langle \mathcal{T}_c\, \epsilon_{\text{dif}(1,3)}(T)\,\epsilon_{\text{dif}(2,4)}(0)\right\rangle_\beta + \ldots\right]. \tag{3.6.27}$$

Note that despite having two difference operators, they are obstructing each other in such a way along the contour that their two-point function does not vanish. Indeed, one finds:

$$\left\langle \mathcal{T}_c\, \epsilon_{\text{dif}(1,3)}(T)\,\epsilon_{\text{dif}(2,4)}(0)\right\rangle_\beta = -\langle[\epsilon(T),\epsilon(0)]\rangle_\beta \sim \frac{1}{2iC}\,e^T\,. \tag{3.6.28}$$

for large T.

Let us explain the above argument in a different way (see [28]). The action (3.6.21) has soft modes that depend exponentially on time:[8]

$$\delta_+\epsilon_I = X^+\,\mathrm{e}^{-t_I(s)}\,, \qquad \delta_-\epsilon_I = -X^-\,\mathrm{e}^{t_I(s)-T}\,. \tag{3.6.29}$$

Along each segment I, these modes have vanishing action. This is related to the unbroken $\text{SL}(2,\mathbb{R})$ symmetry of the Schwarzian action. However, the way these modes contribute to the path integral is slightly more subtle: we can think of $W(T)$ as creating an 'advanced' perturbation $\delta_+\epsilon_I$, which grows towards the past; this excitation is small, of order e^{-T}. However, once sourced on contour segments $I = 2$ and $I = 3$, the effect of this mode will be felt by the early-time operator $V(0)$ on segment $I = 3$ (see figure above). At $t = 0$, the mode $\delta_+\epsilon_I$ has grown to $\mathcal{O}(1)$ and can therefore have a large effect. Similarly, the operator $V(0)$ acts as a source for

[8] We choose convenient (but arbitrary) normalizations such that $\delta_+\epsilon \sim X^+$ at time $t_I = 0$ and $\delta_-\epsilon \sim -X^-$ and $t_I = T$.

the 'retarded' perturbation $\delta_- \epsilon$, which grows large $\mathcal{O}(1)$ as it propagates along the contour to $t = T$. At any point along segment $I = 3$ the interaction between the two modes is $\mathcal{O}(1)$. We summarize this configuration of scramblon modes appropriate for the OTOC computation as follows:

$$\delta\epsilon_I^{\text{otoc}} = (\chi_{I2} + \chi_{I3})\, \delta_+ \epsilon_I + (\chi_{I3} + \chi_{I4})\, \delta_- \epsilon_I \,, \tag{3.6.30}$$

where $\chi_{II} = 1$ and $\chi_{I \neq J} = 0$ with I, J labelling contour segments. We can interpret the first term as capturing the effect of the difference field $\epsilon_{\text{dif}(1,3)}(T)$ and the second term as capturing $\epsilon_{\text{dif}(2,4)}(0)$, which appears in (3.6.27). This is illustrated in the figures above. The reason why the action is not zero can be found in the boundary conditions at the contour turning points, where $\delta\epsilon_I^{\text{otoc}}$ exhibits discontinuities.

We can derive an action, which describes the effective interaction between these modes along the contour (in particular on the segment $I = 3$). Evaluating the quadratic action of Schwarzian fluctuations on $\delta\epsilon_I^{\text{otoc}}$ yields:

$$i S_{\text{quad}}[\delta\epsilon^{\text{otoc}}] = 2iC\, \mathrm{e}^{-T}\, X^+ X^- \,. \tag{3.6.31}$$

The coefficient of this 'eikonal' action, $C\mathrm{e}^{-T}$, is $\mathcal{O}(1)$ at times of order the scrambling time. The path integral over $X^\pm$ must then be performed without resorting to the saddle point approximation.

Evaluating the bilocal operator (3.6.24) on the configuration (3.6.30) yields

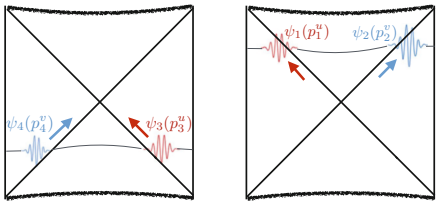
$$\left\langle \mathcal{B}_{1,3}^{\Delta_W}(T,T)\, \mathcal{B}_{2,4}^{\Delta_V}(0,0) \right\rangle_\beta = 1 - \frac{\Delta_V \Delta_W}{\sin^2 \varepsilon} \langle X^+ X^- \rangle_\beta = 1 - \frac{\Delta_V \Delta_W}{2iC\, \sin^2 \varepsilon}\, \mathrm{e}^T \,. \tag{3.6.32}$$

Of course, there are other modes contributing to the OTO path integral. Our claim is merely that $\delta\epsilon_I^{\text{otoc}}$ gives the largest contribution for a large time separation T. Other 'gravitational' (Schwarzian) contributions will be exponentially suppressed at times $t \sim t_\star \equiv \log N$.

3.6.2.3 Gravity and Strings

In preparation for our discussion of scramblons for non-maximal chaos, we briefly review some aspects of the eikonal phase in holography and in string theory. We closely follow [21].

Consider now a more generic case (allowing for higher dimensions) and assume that the theory in question has a semiclassical gravity dual. The thermofield double state corresponds to an eternal black hole geometry and the OTOC (3.6.6) can be represented geometrically as the overlap of two states:



The effect of each of the four operator insertions is approximated by a gravitational null shockwave [19, 36, 37]. On the initial and final time slices we write the corresponding bulk field excitations as wave packets dressed to the boundary via bulk-to-boundary propagators:

$$\begin{aligned} |\Psi_{\text{in}}\rangle &= W_L V_R |\text{TFD}\rangle = \int \mathrm{d}p_3^u \mathrm{d}p_4^v \; \psi_4(p_4^v)\psi_3(p_3^u)\, |p_3^u, p_4^v\rangle_{\text{in}} \,, \\ |\Psi_{\text{out}}\rangle &= V_L W_R |\text{TFD}\rangle = \int \mathrm{d}p_1^u \mathrm{d}p_2^v \; \psi_1(p_1^u)\psi_2(p_2^v)\, |p_1^u, p_2^v\rangle_{\text{out}} \,. \end{aligned} \tag{3.6.33}$$

In Einstein gravity, the wave packets are constructed from bulk-to-boundary propagators, for example:

$$\psi_{1,V}(p_1^u) = \int \mathrm{d}v \; \mathrm{e}^{2ivp_1^u} \, G_{b\partial}(0, v; t_1) \,, \qquad G_{b\partial}(u, v; t) \equiv \langle \phi_V(u, v) V(t)^\dagger \rangle \,. \tag{3.6.34}$$

If the time separation between boundary insertions is large, and hence the relative boost is large, we can approximate $p_1^u \approx p_3^u$ and $p_2^v \approx p_4^v$. In this approximation, assuming the 2-to-2 scattering process is elastic, we furthermore have

$$|p_1^u, p_2^v\rangle_{\text{out}} \approx \mathrm{e}^{i\delta(s,b)} |p_1^u, p_2^v\rangle_{\text{in}} \,, \qquad s = 4p_1^u p_2^v \,, \quad b = |x_1 - x_2| \,. \tag{3.6.35}$$

The OTOC can now be computed as the overlap

$$\text{OTOC} = \langle \Psi_{\text{out}} | \Psi_{\text{in}} \rangle = \int \mathrm{d}p^u \mathrm{d}p^v \; \psi_1(p^u)\psi_3(p^u)\, \psi_2(p^v)\psi_4(p^v)\, \mathrm{e}^{i\delta(s,b)} \,. \tag{3.6.36}$$

In Einstein gravity, the scattering process is dominated by graviton ladder diagrams. The *eikonal phase* $\delta(s, b)$ is proportional to the 2-to-2 scattering amplitude involving a single graviton exchange. It exponentiates in the high energy limit as shown above, and takes the schematic form

$$\delta_{\text{grav}}(s, b) \propto G_N \, s \, \mathrm{e}^{T - \mu b} \,, \tag{3.6.37}$$

where μ is a parameter characterizing the black hole.

We can now use the effective description of scrambling to make an aspect of holography quite manifest. In fact, the gravitational expression (3.6.36) is related to the boundary description derived from the Schwarzian action by a Fourier transform of wave packets with aligned momenta:

$$\begin{aligned}\int \frac{\mathrm{d}p^u}{p^u}\, \psi_{1,V}(p^u)\psi_{3,V}(p^u)\, \mathrm{e}^{-ip^u X^+} &\propto \left[\frac{1}{1-\frac{i}{2\sin\varepsilon}X^+}\right]^{2\Delta_V}, \\ \int \frac{\mathrm{d}p^v}{p^v}\, \psi_{2,W}(p^v)\psi_{4,W}(p^v)\, \mathrm{e}^{-ip^v X^-} &\propto \left[\frac{1}{1-\frac{i}{2\sin\varepsilon}X^-}\right]^{2\Delta_W}.\end{aligned} \tag{3.6.38}$$

The expressions on the right are precisely the two-point functions of V and W in the presence of sources for the scramblon modes excited by the respective other operator. Indeed, these expressions coincide with (3.6.23) when f_I and f_J are suitable non-linear transformations corresponding to the soft modes $\delta_{\pm}\epsilon_I$.

Stringy Effects Corrections to the eikonal phase due to the exchange of a string were studied in [21]. They found that the phase (3.6.37) is replaced by an expression of the following form:

$$\delta_{\text{stringy}} \propto G_N \int \mathrm{d}^{\mathrm{D}-1}p\, \frac{\mathrm{e}^{ip\cdot x_{12}}}{p^2+\mu^2}\left(-i\alpha' s\, \mathrm{e}^T\right)^{J(p)-1}, \tag{3.6.39}$$

with

$$J(p) = 2 - \alpha' \frac{k^2+\mu^2}{2r_0^2} + \dots . \tag{3.6.40}$$

Here, $J(p) < 2$ should be understood as the Regge trajectory of the string parameterized by transverse momentum. The momentum integral is dominated by the *graviton pole* $p^2 = -\mu^2$ when the impact parameter $b = |x_{12}|$ is large. In this case we have the simplification

$$\delta_{\text{stringy}}(b \gg T) \approx \delta_{\text{grav}}\,. \tag{3.6.41}$$

On the other hand, for small impact parameter, stringy effects are important and the integral is no longer dominated by the gravitational contribution. Instead, the saddle point involving $J(p)$ dominates and yields

$$\delta_{\text{stringy}}(b \ll T) \approx G_N\left(-i\alpha' s\, \mathrm{e}^{T-\mu b}\right)^{1-\alpha'\mu^2/2r_0^2} \times \dots \tag{3.6.42}$$

This result was obtained by working to first order in α'. More generally, at finite string length, further modifications might be expected in the bulk computation of the OTOC, (3.6.36): first, the wave functions (bulk-boundary propagators) ψ_i may be modified from their conformally invariant 'gravitational' form. Second, it might be possible that there is an amplitude for *multi-string exchanges* [31]; here, we do not mean the exponentiation of single-string exchanges (which the eikonal phase already captures), but rather the exchange of multiple strings from a single vertex. These effects have not been studied in the bulk theory, but we will see that they appear to play a role indeed in certain boundary models that capture finite coupling effects beyond the Schwarzian approximation.

3.6.2.4 Seeing Strings from Chaos

The effective description of scramblons living and interacting on the OTO contour applies qualitatively almost unchanged in models with non-maximal chaos. We will review below that finite coupling effects in a spatially extended chaotic system can lead to an OTOC that strikingly resembles and generalizes the string theory prediction discussed above.

We will still discuss the example of the SYK model. However, we shall now be interested in the operator spectrum beyond the Schwarzian mode. At finite βJ, the Schwarzian action is no longer parametrically separated, and other modes will contribute to four-point functions.

In order to also discuss spatial dependence, consider a chain of SYK models with nearest-neighbor interactions along the chain. The disorder averaged model turns out to be described by a chain of coupled Liouville-like theories [38]:[9]

$$
\begin{aligned}
i\,S = \frac{C}{4}\sum_{x=0}^{M-1}\int_0^{4T+2\pi} \mathrm{d}s_1\mathrm{d}s_2\,\frac{\mathrm{d}t_I}{\mathrm{d}s_1}\frac{\mathrm{d}t_J}{\mathrm{d}s_2} \\
\times\left[\frac{1}{4}\partial_{t_I} g_{IJ,x}\,\partial_{t_J} g_{IJ,x} - \mathcal{J}_0^2\,\mathrm{e}^{g_{IJ,x}} - \mathcal{J}_1^2\,\mathrm{e}^{\frac{1}{2}(g_{IJ,x}+g_{IJ,x+1})}\right],
\end{aligned}
\tag{3.6.43}
$$

where $g_{IJ,x}(t_I, t_J)$ is bilocal field with a discrete spatial label $x \in \{0, \ldots, M-1\}$ indicating a distance along the chain. The coupling $\mathcal{J}_0^2$ measures an on-site interaction, while $\mathcal{J}_1^2$ parametrizes the interactions between neighboring sites. Two important parameters of the model are $0 \leqslant v < 1$ and $1 < h(p) \leqslant 2$, defined by

$$
\mathcal{J}^2 = \frac{v^2}{4\cos^2\left(\frac{\pi v}{2}\right)}\,, \qquad \frac{h(h-1)}{2} = 1 + \frac{\gamma}{2}\left[\cos(p) - 1\right], \tag{3.6.44}
$$

[9] The parameter $C \equiv \frac{N}{q^2}$ is large, i.e., $1 \ll q^2 \ll N$. It arises from the microscopic description where each site has N Majorana fermions, coupled through Gaußian random variables with q indices.

where p is the spatial momentum along the chain and $\gamma \equiv \mathcal{J}_1^2/(\mathcal{J}_0^2 + \mathcal{J}_1^2)$. The Lyapunov exponent will turn out to be $\kappa(p) = v(h(p) - 1)$ with $\kappa(0) = v$.

The thermal saddle point (for $\beta = 2\pi$) is translation invariant and takes the form

$$e^{g_{IJ,p}(t_I,t_J)} = \frac{\cos^2\left(\frac{\pi v}{2}\right)}{\cos^2\left(\frac{v}{2}\left(\pi - i(t_I - t_J)\right)\right)}\,. \tag{3.6.45}$$

Note that in this model there is no obvious $\mathrm{SL}(2,\mathbb{R})$ symmetry that governs the dynamics. Nevertheless, studying the quadratic action of fluctuations around the above saddle point reveals the existence of zero modes $\delta_\pm g_{IJ,p}$ with exponential time dependence. As before, there is a retarded and an advanced mode, which are excited by the past and future operator insertions, respectively. The OTOC is computed by the following combination:

$$\delta g^{\mathrm{otoc}}_{IJ,p}(t_I, t_J) = X^+(p)\,\delta_+ g_{IJ,p}(t_I, t_J) + X^-(p)\,\delta_- g_{IJ,p}(t_I, t_J)\,, \tag{3.6.46}$$

where

$$\delta_\pm g_{IJ,p}(t_I, t_J) = A_{\pm,IJ}(p)\left[\frac{e^{\pm v((1\mp1)T - t_I - t_J)/2}}{\cos\left(\frac{v}{2}\left(\pi - i(t_I - t_J)\right)\right)}\right]^{h(p)-1}\,, \tag{3.6.47}$$

where $A_{\pm,IJ}(p) = A_{\pm,JI}(p)$ are given by

$$A_{+,IJ}(p) = \begin{cases} e^{-\frac{i\pi}{2}\kappa(p)} & (I,J) = (2,1),(3,1)\,, \\ e^{-\frac{3i\pi}{2}\kappa(p)} & (I,J) = (4,2),(5,2),(4,3),(5,3)\,, \\ 0 & \text{otherwise}\,, \end{cases}$$
$$A_{-,IJ}(p) = \begin{cases} -e^{\frac{3i\pi}{2}\kappa(p)} & (I,J) = (3,1),(4,1),(3,2),(4,2)\,, \\ -e^{\frac{5i\pi}{2}\kappa(p)} & (I,J) = (5,3),(5,4)\,, \\ 0 & \text{otherwise}\,. \end{cases} \tag{3.6.48}$$

These normalization factors are 'locally' arbitrary on each contour segment, but must be chosen in such a way along the contour turning points that $\delta_\pm g_{IJ,p}(s_1, s_2)$ satisfy the KMS conditions

$$g_{5J,p}(4T + 2\pi, s_2) = g_{1J,p}(0, s_2)\,, \qquad g_{I5,p}(s_1, 4T + 2\pi) = g_{I1,p}(s_1, 0)\,. \tag{3.6.49}$$

As in the Schwarzian theory, one finds an effective action for the soft modes $X^{\pm}(p)$:[10]

$$i S_{\text{quad}}[\delta g^{\text{otoc}}_{IJ,p}] = -iC \int_{-\pi}^{\pi} \frac{\mathrm{d}p}{2\pi} \, K\left(\frac{\kappa}{v}\right) \, \cos\left(\frac{\pi\kappa}{2}\right) \, \mathrm{e}^{\frac{i\pi}{2}(\kappa-1)-\kappa T} \, X^{+}(p) X^{-}(-p) \, . \tag{3.6.50}$$

We have highlighted the most important factors, which determine the properties of the propagator $\langle X^{+}(p) X^{-}(-p) \rangle_{\beta}$.

The above formulas are sufficient to compute the OTOC to order $1/C$, which is given by

$$\begin{aligned} &\int \frac{\mathrm{d}p}{2\pi} \, \mathrm{e}^{ipx} \left(\Delta_V \, \delta_{+} g_{42,p}(0,0)\right) \left(\Delta_W \, \delta_{-} g_{31,-p}(T,T)\right) \left\langle X^{+}(p) X^{-}(-p) \right\rangle \\ &\quad = -\frac{\Delta_V \Delta_W}{C} \int \frac{\mathrm{d}p}{2\pi} \, \frac{\mathrm{e}^{ipx+\kappa\left(-\frac{i\pi}{2}+T\right)}}{K\left(\frac{\kappa}{v}\right)\cos\left(\frac{\pi\kappa}{2}\right)} \left[\frac{1}{\cos\left(\frac{\pi v}{2}\right)}\right]^{\frac{2\kappa}{v}} . \end{aligned} \tag{3.6.51}$$

We refer to [31] for details, but only repeat the following points, which are completely analogous to the mechanism that applied for the stringy eikonal scattering phase (3.6.39):

- For $\frac{x}{T}$ is sufficiently small the integral is dominated by a saddle point that extremizes the exponential. The Lyapunov exponent in this case turns out to be v and the OTOC decays exponentially in all spatial directions.
- On the other hand, if $\frac{x}{T}$ is large, then the integral is dominated by the pole at $\kappa = 1$; this pole can be interpreted as due to a "stress tensor" exchange. The resulting Lyapunov exponent is maximal.

In conclusion, the theory of scramblons in the large q SYK chain has a close structural similarity with eikonal scattering in perturbative string theory. In this sense, we see the emergence of a stringy bulk description from the effective theory of chaos. We also point out that there appear to be two competing effects leading to a sub-maximal chaos exponent: on the one hand, the genuinely stringy effects that control the momentum-dependent integral (3.6.51), on the other hand the presence of the v parameter, which "stretches" the thermal circle (e.g., in (3.6.45)). It would be interesting to understand how general this distinction is and if the latter effect has an interpretation in terms of the bulk S-matrix.

[10] For completeness [31]: $K(x) = \frac{1}{2\sqrt{\pi}} \cos(\pi x) \, \Gamma(x+1) \, \Gamma\left(\frac{1}{2} - x\right)$.

Appendix: Schwinger–Keldysh Supersymmetry

It turns out to be rather useful to formulate Schwinger–Keldysh unitarity as a symmetry statement. Clearly, the doubled theory has a topological flavor to it, where the difference operators are sensitive only to 'global' properties. More precisely, we can introduce a *Schwinger–Keldysh BRST cohomology* where the difference operators are BRST-exact under CPT-conjugate charges $\mathcal{Q}_{SK}$ and $\overline{\mathcal{Q}}_{SK}$. For every (say, bosonic) operator $\mathcal{O}$ we introduce a quadruplet consisting of the bosonic operators $\mathcal{O}_{\rm av}$ and $\mathcal{O}_{\rm dif}$, as well as a ghost $\mathcal{O}_G$ and an anti-ghost $\mathcal{O}_{\bar{G}}$. We collect these into a superfield

$$\mathring{\mathcal{O}} = \mathcal{O}_{\rm av} + \theta\,\mathcal{O}_{\bar{G}} + \bar{\theta}\,\mathcal{O}_G + \bar{\theta}\theta\,\mathcal{O}_{\rm dif}\,, \tag{3.6.52}$$

such that translation invariance in the Grassmann-odd directions implements the SK BRST cohomology. For example, $\mathcal{O}_{\rm dif} = \partial_\theta \partial_{\bar{\theta}} \mathring{\mathcal{O}}$ is BRST-exact. This is as desired, but it is a little too naive: the topological nature of $\mathcal{O}_{\rm dif}$ should manifest itself if the operator is future-most within a correlation function. To implement this, we need to fix the behavior of the initial and final states with respect to the BRST charges. One finds that the following rules lead to a consistent picture:

(i) The *final state* (future turning point) is super-translation invariant, i.e., it is annihilated by both BRST charges. This is motivated by the fact that the future turning point in the SK path integral projects onto a maximally entangled state (without any ghosts).

(ii) The *initial state*, being an arbitrary density matrix, generically carries some ghost zero modes. These can, without loss of generality, be parameterized as a background superfield

$$\mathring{\mathcal{O}}_0 \equiv 1 + \theta_0\,\bar{g}_0 + \bar{\theta}_0\,g_0 + \bar{\theta}_0\theta_0\,d_0\,, \tag{3.6.53}$$

which accompanies the initial density matrix in any correlation function, i.e., we effectively replace $\rho_0 \to \mathring{\rho}_0 \equiv \mathring{\mathcal{O}}_0\,\rho_0$.

These conditions can be conveniently formulated in the quadrupled Hilbert space with ghost components: $\mathring{\mathcal{H}} = \mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \mathcal{H}_G \otimes \mathcal{H}_{\bar{G}}$. All operators (including the density matrix) get uplifted to quadruplets in this Hilbert space, and the future turning point is associated with a final state $|\mathring{f}\rangle \in \mathring{\mathcal{H}}$. The conditions above then amount to demanding:

$$\begin{aligned} &(i) \qquad \mathcal{Q}_{SK}|\mathring{f}\rangle = \overline{\mathcal{Q}}_{SK}|\mathring{f}\rangle = 0\,, \\ &(ii) \qquad \mathcal{Q}_{SK}|\mathring{\rho}_0\rangle = g_0|\mathring{\rho}_0\rangle\,, \quad \overline{\mathcal{Q}}_{SK}|\mathring{\rho}_0\rangle = \bar{g}_0|\mathring{\rho}_0\rangle\,. \end{aligned} \tag{3.6.54}$$

Showing consistency of this ansatz amounts to showing that every newly introduced correlator can be consistently fixed in terms of the standard real-time correlators

present in the original SK path integral. This will ensure that there is no net new information and the largest time equation remains upheld. The newly introduced correlators include: (i) those involving any of the SK ghosts, and (ii) those involving the background field $\mathring{\mathcal{O}}_0$. It was shown in [2] that super-translation invariance fixes these consistently will all constraints.

References

1. K.-C. Chou, Z.-B. Su, B.-L. Hao, L. Yu, Equilibrium and nonequilibrium formalisms made unified. Phys.Rept. **118**, 1 (1985)
2. F.M. Haehl, R. Loganayagam, M. Rangamani, Schwinger-Keldysh formalism. Part I: BRST symmetries and superspace. J. High Energy Phys. **06**, 069 (2017). [1610.01940]
3. F.M. Haehl, R. Loganayagam, P. Narayan, M. Rangamani, Classification of out-of-time-order correlators. SciPost Phys. **6**, 001 (2019). [1701.02820]
4. J. Qiao, On the wick rotation of the four-point function in conformal field theories, Ph.D. Thesis, U. PSL (2022). 2209.00285
5. G.'T Hooft, M.J.G. Veltman, DIAGRAMMAR. NATO Sci. Ser. B **4**, 177 (1974)
6. R.P. Feynman, F.L. Vernon, Jr., The theory of a general quantum system interacting with a linear dissipative system. Ann. Phys. **24**, 118 (1963)
7. A. Baidya, C. Jana, R. Loganayagam, A. Rudra, Renormalization in open quantum field theory. Part I. Scalar field theory. J. High Energy Phys. **11**, 204 (2017). [1704.08335]
8. C. Jana, R. Loganayagam, M. Rangamani, Open quantum systems and Schwinger-Keldysh holograms. J. High Energy Phys. **07**, 242 (2020). [2004.02888]
9. R. Loganayagam, M. Rangamani, J. Virrueta, Holographic open quantum systems: toy models and analytic properties of thermal correlators. J. High Energy Phys. **03**, 153 (2023). [2211.07683]
10. F.M. Haehl, R. Loganayagam, P. Narayan, A.A. Nizami, M. Rangamani, Thermal out-of-time-order correlators, KMS relations, and spectral functions. J. High Energy Phys. **12**, 154 (2017). [1706.08956]
11. S. Chaudhuri, C. Chowdhury, R. Loganayagam, Spectral representation of thermal OTO correlators. J. High Energy Phys. **02**, 018 (2019). [1810.03118]
12. P. Glorioso, M. Crossley, H. Liu, A prescription for holographic Schwinger-Keldysh contour in non-equilibrium systems. 1812.08785
13. D.T. Son, A.O. Starinets, Minkowski space correlators in AdS/CFT correspondence: recipe and applications. J. High Energy Phys. **09**, 042 (2002). [hep-th/0205051]
14. K. Skenderis, B.C. van Rees, Real-time gauge/gravity duality: prescription, renormalization and examples. J. High Energy Phys. **05**, 085 (2009). [0812.2909]
15. B.C. van Rees, Real-time gauge/gravity duality and ingoing boundary conditions. Nucl. Phys. B Proc. Suppl. **192–193**, 193 (2009). [0902.4010]
16. R. Loganayagam, G. Martin, An exterior EFT for hawking radiation. 2403.10654
17. S.H. Shenker, D. Stanford, Black holes and the butterfly effect. J. High Energy Phys. **03**, 067 (2014). [1306.0622]
18. S.H. Shenker, D. Stanford, Multiple shocks. J. High Energy Phys. **12**, 046 (2014). [1312.3296]
19. D.A. Roberts, D. Stanford, L. Susskind, Localized shocks. J. High Energy Phys. **03**, 051 (2015). [1409.8180]
20. J. Maldacena, S.H. Shenker, D. Stanford, A bound on chaos. J. High Energy Phys. **08**, 106 (2016). [1503.01409]
21. S.H. Shenker, D. Stanford, Stringy effects in scrambling. J. High Energy Phys. **05**, 132 (2015). [1412.6087]
22. D. Stanford, Many-body chaos at weak coupling. J. High Energy Phys. **10**, 009 (2016). [1512.07687]

23. F.M. Haehl, Y. Zhao, Diagnosing collisions in the interior of a wormhole. Phys. Rev. D **104**, L021901 (2021). [2104.02736]
24. F.M. Haehl, A. Streicher, Y. Zhao, Six-point functions and collisions in the black hole interior. J. High Energy Phys. **08**, 134 (2021). [2105.12755]
25. F.M. Haehl, Y. Zhao, Collisions of localized shocks and quantum circuits. J. High Energy Phys. **09**, 002 (2022). [2202.04661]
26. F.M. Haehl, M. Rozali, Fine grained chaos in AdS_2 Gravity. Phys. Rev. Lett. **120**, 121601 (2018). [1712.04963]
27. T. Anous, F.M. Haehl, On the Virasoro six-point identity block and chaos. J. High Energy Phys. **08**, 002 (2020). [2005.06440]
28. J. Maldacena, D. Stanford, Z. Yang, Conformal symmetry and its breaking in two dimensional nearly anti-de-sitter space. Prog. Theor. Exp. Phys. **2016**, 12C104 (2016). [1606.01857]
29. D. Stanford, Z. Yang, S. Yao, Subleading weingartens. J. High Energy Phys. **02**, 200 (2022). [2107.10252]
30. Y. Gu, A. Kitaev, P. Zhang, A two-way approach to out-of-time-order correlators. J. High Energy Phys. **03**, 133 (2022). [2111.12007]
31. C. Choi, F. M. Haehl, M. Mezei, G. Sárosi, Effective description of sub-maximal chaos: stringy effects for SYK scrambling. J. High Energy Phys. **03**, 142 (2023). [2301.05698]
32. P. Gao, H. Liu, An effective field theory for non-maximal quantum chaos. J. High Energy Phys. **11**, 076 (2023). [2301.05256]
33. S. Sachdev, Bekenstein-Hawking entropy and strange metals. Phys. Rev. X **5**, 041025 (2015). [1506.05111]
34. J. Maldacena, D. Stanford, Remarks on the Sachdev-Ye-Kitaev model. Phys. Rev. D **94**, 106002 (2016). [1604.07818]
35. F.M. Haehl, M. Rozali, *Effective field theory for chaotic CFTs*. J. High Energy Phys. **10**, 118 (2018). [1808.02898]
36. P.C. Aichelburg, R.U. Sexl, On the gravitational field of a massless particle. Gen. Relat. Gravit. **2**, 303 (1971)
37. T. Dray, G.'T Hooft, The gravitational shock wave of a massless particle. Nucl. Phys. B **253**, 173 (1985)
38. C. Choi, M. Mezei, G. Sárosi, Pole skipping away from maximal chaos (2020). 2010.08558

Observables in Expanding Universes

4

Paolo Benincasa and Francisco Vazão

Abstract

Observables in expanding universes are crucial to understand the physics of the early universe. In these lectures, we review some recent progress in understanding their mathematical structure and extract the physics encoded in them. After discussing the most salient features of an expanding background and their consequences for defining an observable, we focus on the so-called Bunch–Davies wavefunctional. We analyze its analytic properties on general grounds and introduce an integral representation for it in perturbation theory for a special class of scalar toy models. We discuss both the diagrammatics associated to the usual Feynman rules and combinatorial rules on the graphs, which generate a representation free of spurious poles. Such combinatorial rules find their origin in the combinatorics of the cosmological polytopes of which we provide a gentle introduction to its definition and its main features. Finally, the combinatorics of the cosmological polytopes turns out to determine the combinatorics of a special class of nestohedra that encode the asymptotic behaviour of the cosmological integrals. We provide a general description of such structures and behaviour, which is of crucial importance to understand the infrared divergences which plague observables in an expanding background.

P. Benincasa (✉)
Max-Planck-Institut für Physik, Werner-Heisenberg-Institut, München, Germany

Instituto Galego de Física de Altas Enerxías IGFAE, Universidade de Santiago de Compostela, Santiago, Spain
e-mail: pablowellinhouse@anche.no

F. Vazão
Max-Planck-Institut für Physik, Werner-Heisenberg-Institut, München, Germany
e-mail: fvvazao@mpp.mpg.de

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

N. Arkani-Hamed et al. (eds.), *Records from the S-Matrix Marathon*,
Lecture Notes in Physics 1041, https://doi.org/10.1007/978-3-031-90352-6_4

Some references include those by D. Anninos [1], P. Benincasa [2, 3], P. Benincasa and F. Vazão [4].

4.1 Generalities and Perturbation Theory

Paolo Benincasa

4.1.1 Motivation

The S-matrix has proven to be well suited to shed light on physics in various setups at appropriate energy scales, and the principles underlying it. In particle-collider experiments, for example, scattering amplitudes between "in" and "out" states provide a good approximation to experimental results. An essential assumption in the theoretical calculations is that the spacetime can be considered approximately flat for distances and times significantly greater than those characteristic of the scattering experiment. However, one of the most striking discoveries of the twentieth century is that our universe is currently undergoing a phase of accelerated expansion, so our spacetime is actually not asymptotically flat. In addition, the universe supposedly had a previous phase of *inflation*—an exponential expansion—during its infancy. Inflation set the seeds for the subsequent evolution and encoded the cosmological structures we observe nowadays, such as temperature fluctuations in the cosmic microwave background. Said differently, the patterns that can be observed in the distribution of galaxies or in the temperature fluctuations in the cosmic microwave background can be traced back to observables at the end of inflation, which essentially provide the initial conditions for the post-inflationary era. Decoding such observables serves two purposes: understanding how the imprints of inflation manifest in structures today, and study the physics of inflation itself. Refer to Fig. 4.1.

One reason for postulating an inflationary phase in the early universe is the need to address the *horizon problem*, i.e., the question of why regions of space that are not causally connected are nevertheless observed to exhibit homogeneity. But yet—and despite its elegance in solving the horizon problem—inflation remains to be better understood: the physics during inflation will not only unveil the phenomena that occurred in the very early universe, but more broadly provide a window into the physics at "ultra high" energies. Given that the Hubble parameter during inflation is estimated to be as large as 10^{14} GeV, it provides an opportunity to explore physics at energy scales that are much higher than those typical of colliders: 10-11 orders of magnitude greater than the energy that can be reached at the Large Hadron Collider at CERN and 5 orders of magnitude smaller than the Planck scale.

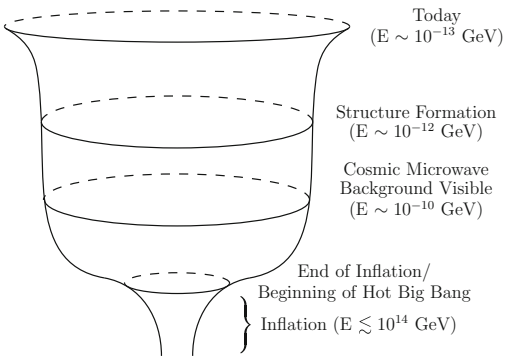
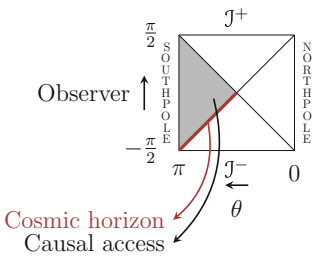


Fig. 4.1 Sketch of the evolution of our universe, where its snapshots relevant to our discussion are emphasised and the related energy scales. The infra-red physics of the large scale structure formation and the cosmic microwave background has its seeds in the initial conditions that result from the inflationary epoque

Both the current and the inflationary expansions can be described in terms of an asymptotically quasi de Sitter (dS) spacetime, with a metric in the conformal global coordinates given by

$$\mathrm{d}s^2_{1+d} = \left(\frac{\ell}{\cos T}\right)^2 [-\mathrm{d}T^2 + \mathrm{d}\theta^2 + \sin^2\theta\, \mathrm{d}\Omega^2_{d-1}]\,, \tag{4.1.1}$$

where $T \in [-\pi/2, \pi/2]$, $\theta \in [0, \pi]$ and d is the number of the spatial dimensions.[1] In order to understand the properties of this metric it is useful to draw its Penrose diagram that depicts an observer in an expanding universe surrounded by a cosmic horizon,



(4.1.2)

In this diagram, each point is a $(d-1)$-dimensional sphere at a fixed value of the coordinate θ, except at the points $\theta = 0$ and $\theta = \pi$ (referred to as the north and south pole, respectively), where the sphere shrinks to a point. A light signal travels at 45° which means that an observer at the south pole, for example, can receive signals from another observer at $\mathcal{I}^-$ (the infinite past) and can send signals to a future observer at $\mathcal{I}^+$ (the infinite future). But that implies that the south-pole observer only

[1] Here, we denote the number of spatial dimensions with d, which is different from the total number of space-time dimensions, $\mathrm{D} = d + 1$, used in other chapters.

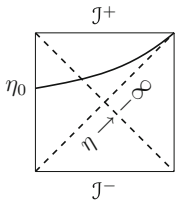
has full causal access to the shaded southern causal diamond, i.e., they can send and receive signals only to points within the shaded region in the diagram.

An observer in an expanding universe is thus surrounded by a *cosmic horizon*: the boundary of the region of spacetime to which they have full causal access. Importantly, the cosmic horizon is observer dependent, so there is not a universal notion of a boundary where our observer can be placed to make observations. In general, the data outside the cosmic horizon do not have a clear physical meaning. In a theory of gravity it is not possible to define local observables, as concentrating too much energy in a finite-sized region of spacetime can cause gravitational backreaction and eventually the formation of black holes, obstructing the ability to perform arbitrarily precise measurements. One can be tempted to consider data at the boundary of the spacetime, as we do in flat space, where we can define an S-matrix. However, this is not possible in a cosmological space-time as, in this case, such data lie outside the observer's horizon.

What partially saves us is that inflation ends at a certain time. Let us consider a constant time-slice in the Poincaré patch, which covers half of dS and is the patch relevant for cosmology. Its metric is

$$\begin{aligned} \mathrm{d}s^2 &= -\mathrm{d}t^2 + \mathrm{e}^{2t/\ell}[r^2\mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\Omega_{d-1}^2] \\ &= (-\ell/\eta)^2[-\mathrm{d}\eta^2 + \mathrm{d}\mathbf{x}^2]\,, \end{aligned} \tag{4.1.3}$$

where $t \in \mathbb{R}$ is the physical time and $\eta \in \mathbb{R}_-$ is the conformal time. This patch corresponds to the upper triangle in the de Sitter Penrose diagram,



(4.1.4)

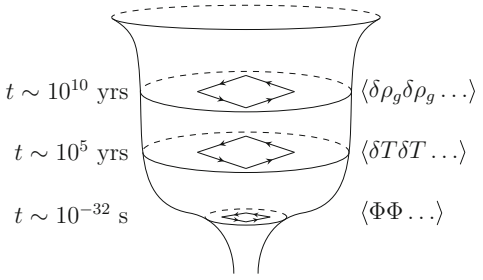
At a given time, $\eta = \eta_0$, an observer can have access to a region determined by η_0, while if they wait long enough, i.e., take $\eta_0 \to 0^-$, all that data becomes data at $\mathcal{I}^+$. This implies that an observer has access to a finite amount of data, and the precision of the measurements are constrained by the size of the data set. Hence, the measurements carried out by a de Sitter observer at any finite time have a classical uncertainty, and, as a consequence, there is a difficulty in measuring a classical quantity even in principle. The key point is:

Any measurement has a classical uncertainty.	(4.1.5)

In other words, one cannot make arbitrarily precise measurements in a de Sitter spacetime.

Audience Question 4.1.1. How does the error on classical uncertainty scale?
Answer: One can, e.g., measure a massive scalar two-point function $\langle\phi\phi\rangle \sim \eta_0^{\Delta}$. Then, the uncertainty on Δ can be quantified as roughly $\sqrt{N}$ where N is the number of modes one has access to. For example, if one takes a wavelength of the size of the universe, there will be a very large uncertainty since we can only make one measurement, but with smaller wavelengths one can make more measurements.

However, assuming that the universe at late enough time becomes infinitely large and flat, then it is possible to define equal-time correlation functions on such spacelike (late time) surfaces,



(4.1.6)

That is, we use this approximation for certain types of modes for which the late slice can be approximated as being at future infinity. One can look for patterns in, for example, the density distribution of galaxies in the temperature fluctuations in the cosmic microwave background. These correlations are the result of the evolution from the initial conditions, which are set by inflation. In other words, the inflationary evolution process encoded in the correlations $\langle\Phi\Phi\cdots\rangle$ at the end of inflation serves as initial conditions for the temperature fluctuation correlations $\langle\delta T\delta T\cdots\rangle$ in the cosmic microwave background and the galaxy distributions, $\langle\delta\rho_g\delta\rho_g\cdots\rangle$. The focus of the following discussion will be on the correlations at the end of inflation $\langle\Phi\Phi\cdots\rangle$ since

(i) as already mentioned, they are the initial conditions that led to the actual observables,
(ii) they are the result of the inflationary evolution and thus they should encode inflationary physics,
(iii) their full-fledged determination might allow us to make precise a notion of holography and, thus, have a formulation of the microscopic theory.

4.1.2 Observables: Correlators, Probability Distributions, Wavefunctions

A general definition of an equal-time correlation function is given as an integral over a certain probability distribution,

$$\langle \Phi(\mathbf{p}_1) \cdots \Phi(\mathbf{p}_n) \rangle = \int \mathcal{D}\Phi P[\Phi] \Phi(\mathbf{p}_1) \cdots \Phi(\mathbf{p}_n), \tag{4.1.7}$$

where $P[\Phi]$ is a probability distribution satisfying $\int \mathcal{D}\Phi P[\Phi] = 1$. We choose to consider the operator $\Phi(\mathbf{p}_1) \cdots \Phi(\mathbf{p}_n)$, but (4.1.7) holds for any quantity $f[\Phi]$ modulo the replacement $\Phi(\mathbf{p}_1) \cdots \Phi(\mathbf{p}_n) \to f[\Phi]$ on both sides.

If we assume that the full system can be described via a wavefunctional $\Psi[\Phi]$, then

$$P[\Phi] = \frac{|\Psi[\Phi]|^2}{\int \mathcal{D}\Phi |\Psi[\Phi]|^2}, \tag{4.1.8}$$

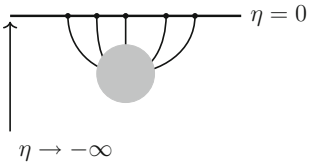
where the wavefunctional is nothing but the transition amplitudes from a vacuum state at past infinity and a state Φ at the quasi-dS space-like future boundary located at $\eta \longrightarrow 0^-$:

$$\Psi[\Phi] = \langle \Phi | \hat{U}(0, -\infty) | 0 \rangle, \tag{4.1.9}$$

with the evolution operator $\hat{U}(0, -\infty)$ defined as

$$\hat{U}(0, -\infty) = \hat{T} \exp\{-i \int_{-\infty}^{0} \mathrm{d}\eta \, H(\eta)\}. \tag{4.1.10}$$

As the correlations are computed on a fixed time-slice—either future infinity at $\eta = 0$ or, in case inflation ends before getting to the future boundary, at some small but finite $\eta = \eta_\circ$ –, they can pictorially be represented as:



(4.1.11)

So far, the definition is pretty general, with the vacuum state yet to be specified. Importantly, the evolution operator is taken to be unitary (and hence $H(\eta)$ to be Hermitian).

Note that expressing the probability distribution in terms of the wavefunctional $\Psi[\Phi]$, the correlator acquires the usual Schwinger–Keldysh form,

$$\langle \Phi \cdots \Phi \rangle = \mathcal{N} \int \mathcal{D}\Phi \, \langle 0|[U(0,-\infty)]^{\dagger}|\Phi\rangle \Phi \cdots \Phi \langle \Phi|U(0,-\infty)|0\rangle \,. \tag{4.1.12}$$

Interestingly, such a formula hints to the fact that the correlation function must inherit some properties of the wavefunctional as well as (at least) a subset of its properties derived from the properties of the wavefunctional. As any average can be computed from a probability distribution expressed in terms of the wavefunctional, the wavefunctional can be considered as a more primitive object. It will be the focus of most of our discussion.

It can be expressed in the more useful form of a path integral,

$$\Psi[\Phi] = \mathcal{N} \int_{\phi(-\infty)=0}^{\phi(0)=\Phi} \mathcal{D}\phi \, \mathrm{e}^{iS[\phi]} \,, \tag{4.1.13}$$

where $S[\phi]$ is an action expressed in therms of the collection of the bulk fields $\phi(\eta, \mathbf{x})$, with boundary conditions taken to be the vacuum in the infinite past, and $\Phi(\mathbf{x})$ at $\mathcal{I}^+$.

The usual choice for a vacuum is the *Bunch–Davies vacuum*, i.e., the one that is exponentially suppressed as $\eta \to -\infty$ while selecting positive-frequency solutions only,

$$\phi(\eta, \mathbf{x}) \xrightarrow[\eta \to -\infty]{} f(\eta) \mathrm{e}^{iE\eta} \,, \tag{4.1.14}$$

with $E = |\mathbf{p}|$ and f is a function which depends on the background (i.e., encodes the cosmology). The factor $\mathrm{e}^{iE\eta}$ shows positive frequencies only and decays exponentially upon suitable regularization; e.g., $\eta \to \eta(1 - i\varepsilon)$ or $E \to E - i\varepsilon$, with $\varepsilon > 0$.

4.1.3 A Comment on the $i\varepsilon$ Prescription

The usual prescription for regulating the wavefunctional as early times are approached is to deform the integration contour around $-\infty$,

$$\hat{U} \to \hat{U}_{\varepsilon} = \hat{T} \exp\left\{ -i \int_{-\infty(1-i\varepsilon)}^{0} \mathrm{d}\eta \, H(\eta) \right\} \,. \tag{4.1.15}$$

However, notice that such a deformation breaks unitarity as $\hat{U}_{\varepsilon}^{\dagger} \neq \hat{U}_{\varepsilon}^{-1}$. A way to obtain convergence while preserving unitarity is via deforming the Hamiltonian rather than the integration contour, while preserving its Hermiticity [5]:

$$H(\eta) \to \mathrm{e}^{\varepsilon\eta} H(\eta) \,. \tag{4.1.16}$$

More generally, one can give a small negative imaginary part to the energies. One can take different a ε for each energy, $E_j \to E_j - i\varepsilon_j$, $\varepsilon_j > 0$ [6]. The above deformation of the Hamiltonian is just a special case of such a deformation.

Audience Question 4.1.2. What happens in the limit as $\varepsilon \to 0$?
Answer: Unitarity is restored in the strict limit $\varepsilon \to 0$. Nevertheless, the prescription above is important for manifestly recovering the flat-space limit of the cosmological cutting rules, as well as non-perturbatively.

4.1.4 Perturbative Rules for the Bunch–Davies Wavefunction

The path integral formulation of the wavefunctional allows us to easily derive the Feynman rules in perturbation theory. In particular, if we split ϕ into a classical free part $\phi_{\rm cl}$ and its fluctuations φ,

$$\phi \to \phi_{\rm cl} + \varphi\,, \tag{4.1.17}$$

then,

$$\psi[\Phi] = \mathcal{N}\int_{\phi(-\infty)=0}^{\phi(0)=\Phi} \mathcal{D}\phi\, \mathrm{e}^{iS[\phi]} = \mathrm{e}^{iS_2[\phi_{\rm cl}]}\mathcal{N}'\int_{\varphi(-\infty)=0}^{\varphi(0)=0} \mathcal{D}\varphi\, \mathrm{e}^{iS_2[\varphi]}\mathrm{e}^{iS_{\rm int}[\phi_{\rm cl},\varphi]}\,, \tag{4.1.18}$$

where S_2 indicates the free action—the part that is evaluated at the classical solution while the integral is just over the fluctuations. The factors $\mathcal{N}$ and $\mathcal{N}'$ are normalizations. Importantly, the boundary condition of the fluctuations made them vanish both at the infinite past and at $\mathcal{I}^+$. This is easy to understand since

$$\phi(\eta=0)=\Phi \quad \Rightarrow \quad \phi_{\rm cl}(\eta=0)=\Phi\,, \qquad \varphi(\eta=0)=0\,, \tag{4.1.19}$$

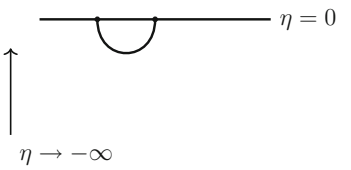
$$\phi(\eta\to-\infty)=0 \quad \Rightarrow \quad \phi_{\rm cl}(\eta\to-\infty)=0\,, \qquad \varphi(\eta\to-\infty)=0\,, \tag{4.1.20}$$

expanding in powers of $S_{\rm int}$, the path-integral expression can be organized as

$$\Psi[\Phi] = \exp\left\{-\prod_{j=1}^{2}\left[\int \frac{\mathrm{d}^d\mathbf{p}_j}{(2\pi)^d}\Phi(\mathbf{p}_j)\right]\psi_2(\mathbf{p}_1,\mathbf{p}_2)\right\} \times\left\{1+\sum_{n>0}\int\prod_{k=1}^{n}\left[\frac{\mathrm{d}^d\mathbf{p}_k}{(2\pi)^d}\Phi(\mathbf{p}_k)\right]\psi_n(\mathbf{p}_1,\ldots,\mathbf{p}_n)\right\}\,, \tag{4.1.21}$$

with the "wavefunction coefficients" $\psi_n(\mathbf{p}_1, ..., \mathbf{p}_n)$, which can be expressed in terms of Feynman graphs:

(i) ψ_2 appearing in the factorized exponential is just the free two-point wavefunction and arises from the boundary term of the free on-shell action. Notice that the Feynman graphs are characterized by having their external states ending on a horizontal line that represents the future spacelike boundary $\mathcal{I}^+$ at $\eta = 0$; the two-point wavefunction, for example, has support on $\delta^{(d)}(\mathbf{p}_1 + \mathbf{p}_2)$.



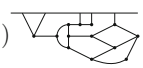
(4.1.22)

(ii) $\psi_n(\mathbf{p}_1, \ldots, \mathbf{p}_n)$ are the n-point wavefunction coefficients, which show both connected and disconnected components, and that receive contributions—in principle—at all loops:

$$\psi_n = \sum_{\mathrm{L}\geqslant 0} \left[\prod_{\ell=0}^{\mathrm{L}} \int \frac{\mathrm{d}^d \boldsymbol{\ell}_e}{(2\pi)^d} \Psi^{(\ell)}(\mathbf{p}_1, \ldots, \mathbf{p}_n, \boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_{\mathrm{L}}) \right], \tag{4.1.23}$$

with

$$\Psi^{(\ell)}(\mathbf{p}_1, \ldots, \mathbf{p}_n, \boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_{\mathrm{L}})\big|_{\text{ connected}} = \delta^{(d)}(\mathbf{p}_1 + \ldots + \mathbf{p}_n)$$



$$= \delta^{(d)}(\mathbf{p}_1 + \ldots + \mathbf{p}_n) \int_{-\infty}^{0} \prod_{s\in\mathcal{V}} \mathrm{d}\eta_s V(\eta_s) \left[\prod_{j\in S} \phi_0(-E_j \eta_s) \right] \prod_{e\in\mathcal{E}} G(y_e; \eta_{s_e} \eta_{s'_e}), \tag{4.1.24}$$

where $\mathcal{V}$ is the set of sites of the graph, while $\mathcal{E}$ is the set of edges. Furthermore,

- $V(\eta_s)$ is the vertex function at the site s, which can depend on the warp factor of the metric—here and in what follows we consider processes that occur in conformally flat backgrounds,

$$\mathrm{d}s^2_{1+d} = a^2(\eta)\big[-\mathrm{d}\eta^2 + \mathrm{d}\mathbf{x}^2 \big], \qquad \eta \in \mathbb{R}_- . \tag{4.1.25}$$

  The vertex function can further depend on scalar products of momenta (in the case of derivative interactions—for simplicity we will restrict to the case of polynomial interactions, unless otherwise specified).

- $\phi_0(-E_j\eta_s)$ is the classical solution stripped out of the boundary value $\Phi(\mathbf{p})$, e.g.,

$$\phi_{\rm cl}(\mathbf{p},\eta) = \Phi(\mathbf{p})\phi_0(-E_j\eta_s). \tag{4.1.26}$$

  Hence, $\phi_0(-E_j\eta_s)$ is taken to normalize to 1, that is, $\phi_0(-E_j\eta = 0) = 1$.
- $G(y_e;\eta_{s_e},\eta_{s'_e})$ is the propagator that connects two sites s_e and s'_e, with momentum $\mathbf{p}_e$ flowing through it, and $y_e \equiv |\mathbf{p}_e|$.

Because the fluctuations have to vanish at the boundary $\eta = 0$, the propagator has a three-term structure: two of them are the usual retarded and advanced contributions while the third one is purely a boundary term that manifestly breaks time-translation invariance:

$$\begin{aligned} G(y_e;\eta_{s_e},\eta_{s'_e}) = \frac{1}{2\,\mathrm{Re}\left\{\Psi_2(y_e)\right\}}\Big[&\bar{\phi}_0(-y_e\eta_{s_e})\phi_0(-y_e\eta_{s'_e})\Theta(\eta_{s_e}-\eta_{s'_e}) \\ &+\phi_0(-y_e\eta_{s_e})\phi_0(-y_e\eta_{s'_e})\Theta(\eta_{s'_e}-\eta_{s_e}) \\ &-\phi_0(-y_e\eta_{s_e})\phi_0(-y_e\eta_{s'_e})\Big], \end{aligned} \tag{4.1.27}$$

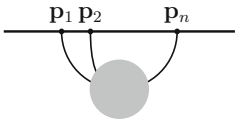
where $\mathrm{Re}\left\{\Psi_2(y_e)\right\}$ is the real part of the free two-point wavefunction.

4.1.5 The Wavefunction and Its Analytic Properties

Despite the fact that so far the discussion has been very general, it is nevertheless possible to extract important information on the expected analytic structure of the wavefunction.

First, consider a graph $\mathcal{G}$ that contributes to $\psi_n^{(\ell)}$. Following the Feynman rules discussed earlier, its contribution $\psi_{\mathcal{G}}$ can be written as

$$\psi_{\mathcal{G}} = \int_{-\infty}^{0}\prod_{s\in\mathcal{V}}[\mathrm{d}\eta_s V(\eta_s)]\prod_{j\in s}\phi_0(-E_j\eta_s)\prod_{e\in\mathcal{E}}G(y_e;\eta_{s_e},\eta_{s'_e}) \tag{4.1.28}$$

$=$ 

(4.1.29)

Let $\bar{\eta} = \eta_1 + \ldots + \eta_{n_s}$ be the "center of mass" time (with $n_s = \dim\{\mathcal{V}\}$) and $\eta_{s_e s'_e} = \frac{1}{2}(\eta_{s_e} - \eta_{s'_e})$.

Then, as $\bar{\eta}\to-\infty$, $\phi_0(-E_j\eta_s)\to f(\eta_s)\mathrm{e}^{iE_j\cdot\eta_s}$ because of the Bunch–Davies condition, while $G(y_e;\eta_{s_e},\eta_{s'e})\to G_F(y_e;\eta_{s_e s'e})\equiv G_R(y_e;\eta_{s_e s'_e})+G_A(y_e;\eta_{s_e s'_e})$ where G_F, G_R, and G_A correspond to the Feynman, retarded, and advanced

propagators, respectively. That is, the propagator reduces to the Feynman propagator in this limit. Hence,

$$\psi_{\mathcal{G}} \to \int_{-\infty} \mathrm{d}\bar{\eta}\, \tilde{V}(\bar{\eta}) \mathrm{e}^{i(E_1+\ldots+E_n)\bar{\eta}} \int_{-\infty}^{\infty} \prod_{e\in\mathcal{E}} \mathrm{d}\eta_{s_e s_e'} \mathrm{e}^{iE_{s_e s_e'}\eta_{s_e s_e'}} \prod_{e\in\mathcal{E}} G_F(y_e; \eta_{s_e s_e'}) \,. \tag{4.1.30}$$

Note that at a generic point in kinematic space, the integral is exponentially suppressed by $\mathrm{e}^{i(E_1+\ldots+E_n)\bar{\eta}}$ upon regularization. However, for $E_1 + \ldots + E_n = 0$ the integral becomes singular with the type of singularity which depends on the function $\tilde{V}(\eta)$ and its coefficient given by the remaining integral which, for states that have a flat-space counterpart, is given by the high-energy limit of the graph contribution to the flat-space amplitude. Importantly, the locus $E_1 + \ldots + E_n = 0$ can be non-trivially reached if the energies are analytically continued such that a subset of them becomes negative—such an analytic continuation maps these states to become "out" states. Furthermore, as $\bar{\eta} \to -\infty$, the spacelike boundary becomes effectively infinitely far away, which can be seen from the fact that the propagator G reduces to the Feynman propagator. This creates the setup for a flat-space scattering, with $E_{\text{tot}} = E_1 + \ldots + E_n = 0$ only having support on the total energy-conservation pole.

Note that the behavior just described is generic and does not depend on whether the states are massless or not: in all cases the mode functions reduce to the same exponential as a consequence of the Bunch–Davies condition in the far past. For example, in de Sitter space, massive states belong to its principal series and are characterized by a mode function given by a Hankel function with imaginary order parameter:

$$\phi_0(-E\eta) = \sqrt{-E\eta}\, H^{(2)}_{i\lambda}(-E\eta) \,. \tag{4.1.31}$$

However, as $-E\eta \to \infty$, we get

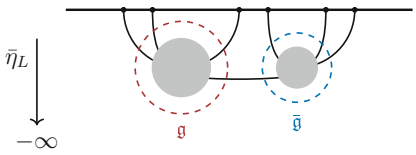
$$\phi_0(-E\eta) \xrightarrow{\sim} \mathrm{e}^{iE\eta} \,. \tag{4.1.32}$$

In this sense, one obtains the high energy limit of a flat-space amplitude.

As a final comment, for states that do not have a flat-space counterpart, the analysis holds, with the important difference that now its coefficient no longer can be interpreted as the flat-space amplitude.

In summary, the Bunch–Davies wavefunction is singular on the locus $E_{\mathcal{G}} \equiv E_1 + \ldots + E_n = 0$ that lives outside the physical region and thus can be reached just upon analytic continuation.

The very same analysis can be carried out by considering the center-of-mass time associated to a given sub-process, identified by a subgraph $\mathfrak{g}$:



(4.1.33)

In this case the singularity is located at

$$E_{\mathfrak{g}} \equiv \sum_{s\in\mathcal{V}_{\mathfrak{g}}} \sum_{j\in s} E_j + \sum_{c\in\mathcal{E}^{\text{ext}}_{\mathfrak{g}}} y_e = 0\,, \tag{4.1.34}$$

where $\mathcal{V}_{\mathfrak{g}}$ is the set of sites in $\mathfrak{g}$, while $\mathcal{E}^{\text{ext}}_{\mathfrak{g}}$ is the set of edges departing from $\mathfrak{g}$. On this hyperplane, the wavefunction coefficient factorizes into a lower-point scattering amplitude and a linear combination of lower point wavefunctions:

$$\tilde{\psi}_{\mathcal{G}} \xrightarrow{E_{\mathfrak{g}}\to 0} A_{\mathfrak{g}} \sum_{\sigma_{\ell}=\pm 1} \frac{1}{2\,\text{Re}\left\{\Psi_2(y_{\ell})\right\}} \Psi_{\bar{\mathfrak{g}}}(\vec{\ell}(\sigma_{\ell}), R)\,, \tag{4.1.35}$$

where $A_{\mathfrak{g}}$ is the contribution to the flat-space amplitude associated to the subgraph $\mathfrak{g}$, while $\Psi_{\bar{\mathfrak{g}}}(\vec{\ell}, R)$ is the wavefunction associated with $\bar{\mathfrak{g}}$ whose external energies contain the energies of the edges $\vec{\ell}$ which connect it to $\mathfrak{g}$ each of which is taken with the sign σ_{ℓ}. This can be understood thinking that in this limit just one of the time-ordered terms in the propagator $\mathfrak{g}$ survives, e.g.,

$$G(y_{\ell};\eta_{s_{\ell}},\eta_{s'_{\ell}}) \to G_{\text{ret}} - G_{\partial} = \frac{1}{2\,\text{Re}\left\{\Psi_2(y_{\ell})\right\}} \phi(-y\eta_{s'_{\ell}}) \left[\bar{\phi}_0(-y\eta_{s_{\ell}}) - \phi_0(-y\eta_{s_{\ell}})\right], \tag{4.1.36}$$

giving rise to the sum in the right-hand side of (4.1.35).

In summary, the wavefunction shows singularities in correspondence with the vanishing loci of the energies associated to its sub-processes, $\{E_{\mathfrak{g}} = 0, \quad \mathfrak{g} \subseteq \mathcal{G}\}$. The coefficient of each of these singularities is the factorization into a lower-point amplitude and a linear combination of lower-point wavefunctions.

Hence, given a graph $\mathcal{G}$, there is a one-to-one correspondence between the singularities of $\psi_{\mathcal{G}}$ and the subgraphs $\{\mathfrak{g} \subseteq \mathcal{G}\}$.

4.1.6 Cosmological Integrals

Let us consider a large class of scalar toy models consisting of polynomial interactions with time-dependent couplings:

$$S[\phi] = -\int_{-\infty}^{0} \mathrm{d}\eta \int \mathrm{d}^d\mathbf{x} \left[\frac{1}{2}(\partial\phi)^2 - \sum_{k\geqslant 3} \frac{\lambda_k(\eta)}{k!}\phi^k \right] . \tag{4.1.37}$$

This class contains as a special case conformally-coupled scalars in FRW-cosmologies, provided that $\lambda_k(\eta)$ is identified to be

$$\lambda_k(\eta) := \lambda\,[a(\eta)]^{2-\frac{(k-2)(d-1)}{2}} , \tag{4.1.38}$$

where $a(\eta)$ is the warp factor of the FRW metric $\mathrm{d}s^2 = a^2(\eta)[-\mathrm{d}\eta^2 + \mathrm{d}\mathbf{x}^2]$.

Furthermore, the conformally-coupled scalar can be mapped to other states via a differential operator. Then, given a graph $\mathcal{G}$, the wavefunction coefficient associated to it is given by

$$\tilde{\psi}_{\mathcal{G}} = i^{n_s} \int_{-\infty}^{0} \prod_{s\in\mathcal{V}} \mathrm{d}\eta_s \lambda_k(\eta_s) \mathrm{e}^{iX_s\eta_s} \prod_{e\in\mathcal{E}} G(y_e;\, \eta_{s_e},\, \eta_{s'e}) \,, \tag{4.1.39}$$

where $X_s = \sum_{j\in s} E_j$ is the total energy of the states in the graph $\mathcal{G}$, and n_s is the number of external states in $\mathcal{G}$. The mode functions are simple exponentials, and thus the propagators are,

$$G(y_e;\, \eta_{s_e},\, \eta_{s'e}) = \frac{1}{2y_e}\Big[\mathrm{e}^{-iy_e(\eta_{s_e}-\eta_{s'_e})}\Theta(\eta_{s_e}-\eta_{s'_e}) + \mathrm{e}^{iy_e(\eta_{s_e}-\eta_{s'e})}\Theta(\eta_{s'e}-\eta_{s_e}) - \mathrm{e}^{iy_e(\eta_{s_e}+\eta_{s'_e})}\Big] . \tag{4.1.40}$$

Finally, the spatial momentum-conserving delta function has been stripped off.

In order to perform the time integrals over η, it naively seems to be necessary to specify $\lambda(\eta)$ at this stage, which encodes the details of the cosmology. However, we can still extract more information while staying agnostic to the details of the cosmology. To that end, we consider the following integral expression for $\lambda(\eta)$:

$$\lambda(\eta) := i\int_{-\infty}^{\infty} \mathrm{d}z\, \mathrm{e}^{iz\eta}\tilde{\lambda}(z) = \int_{0}^{\infty} \mathrm{d}z\, \mathrm{e}^{iz\eta}\tilde{\lambda}(z), \tag{4.1.41}$$

with the second equality coming from the fact that $\lambda(\eta)$ has support only on $\Theta(-\eta)$ as $\eta \in \mathbb{R}_-$. Hence,

$$\tilde{\psi}_{\mathcal{G}} = \int_{0}^{\infty} \prod_{s\in\mathcal{V}} \mathrm{d}z_s\, \tilde{\lambda}(z_s) i^{n_s} \int_{-\infty}^{0} \prod_{s\in\mathcal{V}} d\eta_s\, e^{i(X_s+z_s)\eta_s} \prod_{e\in\mathcal{E}} G(y_e;\, \eta_{s_e},\, \eta_{s'_e}) \tag{4.1.42}$$

$$= \prod_{s\in\mathcal{V}} \int_{X_s}^{\infty} \mathrm{d}x_s\, \tilde{\lambda}(x_s - X_s) i^{n_s} \int_{-\infty}^{0} \prod_{s\in\mathcal{V}} \mathrm{d}\eta_s \mathrm{e}^{ix_s\eta_s} \prod_{e\in\mathcal{E}} G(y_e; \eta_{s_e}, \eta_{s'_e}) \tag{4.1.43}$$

$$= \prod_{s\in\mathcal{V}} \int_{X_s}^{\infty} \mathrm{d}x_s \tilde{\lambda}(x_s - X_s) \psi_{\mathcal{G}}(x, y), \tag{4.1.44}$$

where the second line comes from the simple change of variables $x_s = X_s + z_s$. Furthermore, recall that

$$\psi_{\mathcal{G}}(x, y) \equiv i^{n_s} \int_{-\infty}^{0} \prod_{s\in\mathcal{V}} \mathrm{d}\eta_s \mathrm{e}^{ix_s\eta_s} \prod_{e\in\mathcal{E}} G(y_e; \eta_{s_e}, \eta_{s'_e}), \tag{4.1.45}$$

is nothing but the flat-space wavefunction, with $x \equiv \{x_s\}_{s\in\mathcal{V}}$ and $\{y_e\}_{e\in\mathcal{E}}$ parametrising the flat-space kinematic space.

Hence, the formula

$$\tilde{\psi}_{\mathcal{G}}(X, y) = \prod_{s\in\mathcal{V}} \int_{X_s}^{\infty} \mathrm{d}x_s\, \tilde{\lambda}(x_s - X_s) \psi_{\mathcal{G}}(x, y)\,, \tag{4.1.46}$$

can be seen as a map between the flat-space wavefunction and the FRW one, with the details of the cosmology encoded in the integration measure $\tilde{\lambda}(x_s - X_s)$. This formula makes explicit that part of the structure of $\tilde{\psi}_{\mathcal{G}}(X, y)$ is determined by $\psi_{\mathcal{G}}(x, y)$—in particular the loci where a subset of energies vanishes, $\{E_{\mathfrak{g}}(x, y) = 0,\ \mathfrak{g} \subseteq \mathcal{G}\}$, are mapped into the loci $\{E_{\mathfrak{g}}(X, y) = 0,\ \mathfrak{g} \subseteq \mathcal{G}\}$. The full singularity structure of $\tilde{\psi}_{\mathcal{G}}$ is further constrained by the way in which the measure of integration $\tilde{\lambda}$, the integrand $\psi_{\mathcal{G}}(x, y)$, and the integration contour over x_s relate to each other.

For power-law cosmologies with $a(\eta) = \left(-\frac{\ell}{\eta}\right)^{\gamma}$ ($\mathrm{Re}(\gamma) > 0$), the map is simply a Mellin transform

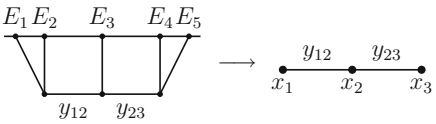
$$\tilde{\psi}_{\mathcal{G}}(X, y) = \frac{i^{\alpha}}{\Gamma(\alpha)} \ell^{\alpha} \prod_{s\in\mathcal{V}} \int_{X_s}^{\infty} \mathrm{d}x_s (x_s - X_s)^{\alpha-1} \psi_{\mathcal{G}}(x, y) \tag{4.1.47}$$

$$= \frac{i^{\alpha}}{\Gamma(\alpha)} \ell^{\alpha} \int_{0}^{\infty} \prod_{s\in\mathcal{V}} \frac{\mathrm{d}x_s}{x_s} x_s^{\alpha}\, \psi_{\mathcal{G}}(x + X, y)\,. \tag{4.1.48}$$

As is customary in the literature on scattering amplitudes, it is possible to study the integrand $\psi_{\mathcal{G}}(x, y)$ and later extend the information on the integrated function. We stress once more that the integrand is the same for each cosmology: its structure will affect the one of $\tilde{\psi}_{\mathcal{G}}(X, y)$ for any cosmology.

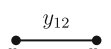
First, note that the wavefunction coefficient $\psi_{\mathcal{G}}(x, y)$ associated with a graph $\mathcal{G}$ depends only on the labels associated with the sites of $\mathcal{G}$, $\{x_s, s \in \mathcal{V}\}$, and on its edges $\{y_e, e \in \mathcal{E}\}$. Hence, rather than considering the full graph, it is possible to

define a reduced graph that associates a weight to each site, e.g.,



(4.1.49)

with $E_j = |\mathbf{p}_j|$, $y_{12} = |\mathbf{p}_1+\mathbf{p}_2|$ and $y_{23} = |\mathbf{p}_2+\mathbf{p}_3|$. The reduced graph on the right-hand side is simply obtained by suppressing the external legs and instead weighting each site s with the external energies $x_j = \sum_{j\in s} |\mathbf{p}_s|$. In particular, the weighted reduced graph on the right-hand side of (4.1.49) has $x_1 = E_1 + E_2$, $x_2 = E_3$ and $x_3 = E_4 + E_5$. The integration over $\{x_s, s \in \mathcal{V}\}$ then maps the flat-space kinematics onto the FRW one. Then, the reduced weighted graph represents the time integration in flat space. Direct integration returns an expression with 3^{n_e} terms ($n_e = \dim\{\mathcal{E}\}$) with spurious poles at $\{y_e = 0,\ e \in \mathcal{E}\}$ because of the 3-term structure of G, e.g.,

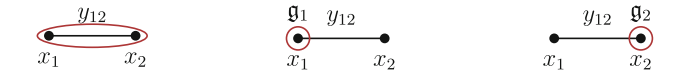


$$
\begin{aligned}
&= i^2 \int_{-\infty}^{0} \mathrm{d}\eta_1 \mathrm{e}^{ix_1\eta_1} \int_{-\infty}^{0} \mathrm{d}\eta_2 \mathrm{e}^{ix_2\eta_2} G(y_{12};\eta_1,\eta_2) \\
&= \underbrace{\frac{1}{2y_{12}(x_1+x_2)(x_2+y_{12})}}_{\text{retarded contribution}} + \underbrace{\frac{1}{2y_{12}(x_1+x_2)(x_1+y_{12})}}_{\text{advanced contribution}} - \underbrace{\frac{1}{2y_{12}(x_1+y_{12})(y_{12}+x_2)}}_{\text{boundary term}} \\
&= \frac{1}{(x_1+x_2)(x_1+y)(y+x_2)}\,.
\end{aligned}
\tag{4.1.50}
$$

The last line explicitly shows that

(i) $y_{12} = 0$ is a spurious pole,

(ii) the graph has only singularities corresponding to subgraphs, as predicted by the general argument discussed earlier



(4.1.51)

with the general formula $E_{\mathfrak{g}} = \sum_{s\in\mathcal{V}_s} x_s + \sum_{e\in\mathcal{E}^{\text{ext}}_{\mathfrak{g}}} y_e$, where $\mathcal{V}_s$ is the set of sites in $\mathfrak{g}$, while $\mathcal{E}^{\text{ext}}_{\mathfrak{g}}$ is the set of edges departing from $\mathfrak{g}$,

(iii) the expression obtained via Feynman rules is redundant and can be simplified in one which contains just physical singularities.

A general question is whether such a formula can be obtained, bypassing the Feynman rules, and whether there is a mathematical structure behind it.

Let us consider a general graph $\mathcal{G}$ and the time integral associated to it

$$\psi_{\mathcal{G}} = i^{n_s} \int_{-\infty}^{0} \prod_{s\in\mathcal{V}} \Big[\mathrm{d}\eta_s \mathrm{e}^{ix_s\eta_s} \Big] \prod_{e\in\mathcal{E}} G(y_e; \eta_{s_e}, \eta_{s'_e}) \,. \tag{4.1.52}$$

Now let us consider the very same integral but with the total time translation generator acting on the full integrand—this integral is zero because of the boundary conditions:

$$0 = i^{n_s} \int_{-\infty}^{0} \prod_{s\in\mathcal{V}} [\mathrm{d}\eta_s]\, \hat{\Delta} \left\{ \prod_{s\in\mathcal{V}} \mathrm{e}^{ix_s\eta_s} \prod_{e\in\mathcal{E}} G(y_e; \eta_{s_e}, \eta_{s'_e}) \right\} , \tag{4.1.53}$$

with $\hat{\Delta} = -i \sum_{s\in\mathcal{V}} \partial_{\eta_s}$. However, we can make $\hat{\Delta}$ act on the external states $\mathrm{e}^{ix_s\eta_s}$ and then on the propagators $G(y_e; \eta_{s_e}, \eta_{s'_e})$,

$$0 = i^{n_s} \int_{-\infty}^{0} \prod_{s\in\mathcal{V}} [\mathrm{d}\eta_s]\, \hat{\Delta} \Big[\prod_{s\in\mathcal{V}} \mathrm{e}^{ix_s\eta_s} \Big] \prod_{e\in\mathcal{E}} G(y_e; \eta_{s_e}, \eta_{s'_e}) \tag{4.1.54}$$

$$+ i^{n_s} \int_{-\infty}^{0} \prod_{s\in\mathcal{V}} \Big[\mathrm{d}\eta_s \mathrm{e}^{ix_s\eta_s} \Big] \hat{\Delta} \Big[\prod_{e\in\mathcal{E}} G(y_e; \eta_{s_e}, \eta_{s'_e}) \Big] \tag{4.1.55}$$

$$= \Big(\sum_{s\in\mathcal{V}x_s} x_s \Big) \psi_{\mathcal{G}} - \sum_{e\in\mathcal{E}} \psi_{\mathcal{G}\setminus\{e\}}(x_j^{(e)}, g) \,, \tag{4.1.56}$$

where $\mathcal{G}\setminus\{e\}$ represents the graph obtained by deleting the edge, while $x_s^{(e)} := x_s + \sum_{\{e\}\cap\mathcal{E}_s} y_e$, $\mathcal{E}_s$ being the set of edges departing from s. The relation above provides a recursion relation. Note that $\mathcal{G}\setminus\{e\}$ can be either disconnected or connected. Separating these two classes of contributions then results in

$$\Big(\sum_{s\in\mathcal{V}x_s} x_s \Big) \psi_{\mathcal{G}} = \sum_{e\in\mathcal{E}} \underbrace{\quad}_{\text{disconnected terms}} + \sum_{e\in\mathcal{E}} \underbrace{\quad}_{\text{connected terms}} \,, \tag{4.1.57}$$

with the dashed line indicating the deleted edge. Note that the term $\big(\sum_{s\in\mathcal{V}} x_s\big)\psi_{\mathcal{G}}(x, y)$ has been obtained because

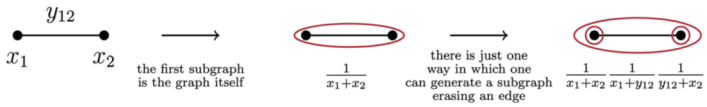
$$\hat{\Delta} \prod_{s\in\mathcal{V}} \mathrm{e}^{ix_s\eta_s} = \Big(\sum_{s\in\mathcal{V}} x_s \Big) \prod_{s\in\mathcal{V}} \mathrm{e}^{ix_s\eta_s} \,, \tag{4.1.58}$$

the term $\sum_{e\in\mathcal{E}} \psi_{\mathcal{G}\setminus\{e\}}(x_j^{(e)}, y)$ instead uses the relation $\hat{\Delta} G = -\mathrm{e}^{iy_e(\eta_{s_e}+\eta_{s'_e})}$.

The resulting recursion relation can be translated into simple graphical rules, by iteratively considering all possible subgraphs and summing all possible ways

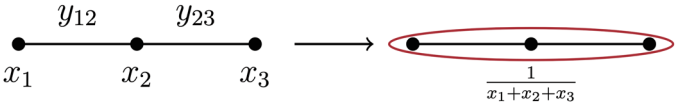
in which subgraphs can be obtained by erasing one edge, while associated to each subgraph the inverse of its total energy $1/E_{\mathfrak{g}}$, $E_{\mathfrak{g}} := \sum_{s\in\mathcal{V}_{\mathfrak{g}}} x_s + \sum_{e\in\mathcal{E}^{\text{ext}}_{\mathfrak{g}}} y_e$.

Example 4.1.1. We first look at the simplest example with only two sites. The first subgraph is simply the graph itself:



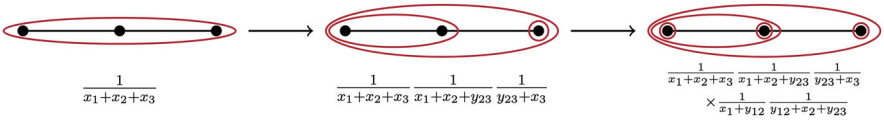
(4.1.59)

Example 4.1.2. Next, we look at an example with three sites. The first encircling is the one of all the sites:



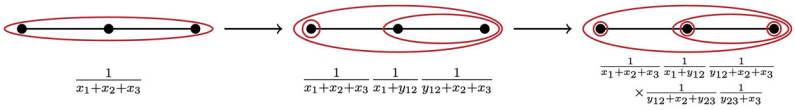
(4.1.60)

There are now two different ways of generating subgraphs by deleting an edge. In both cases, there is only one way to generate further subgraphs by deleting one edge. The first possibility is,



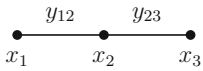
(4.1.61)

while the second one is



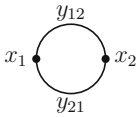
(4.1.62)

The final result is the sum of such contributions:



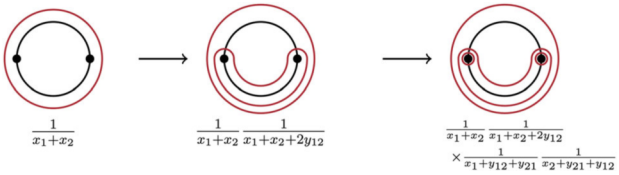
$$= \frac{1}{(x_1+x_2+x_3)(x_1+y_{12})(y_{12}+x_2+y_{13})(y_{23}+x_3)} \times \left[\frac{1}{x_1+x_2+y_{23}} + \frac{1}{y_{12}+x_2+x_3}\right]. \tag{4.1.63}$$

Example 4.1.3. As a last example, let us consider the 2-site one-loop graph,



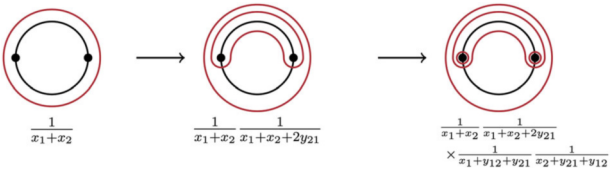
(4.1.64)

Again, we have two different ways of deleting edges. The first possibility is



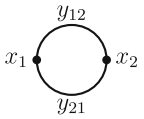
(4.1.65)

and the second one is



(4.1.66)

The final result for the one-loop graph is therefore the sum of these two contributions:



$$= \frac{1}{(x_1+x_2)(x_1+y_{12}+y_{21})(x_2+y_{12}+y_{21})} \times \left[\frac{1}{x_1+x_2+2y_{12}} + \frac{1}{x_2+x_2+2y_{21}}\right]. \tag{4.1.67}$$

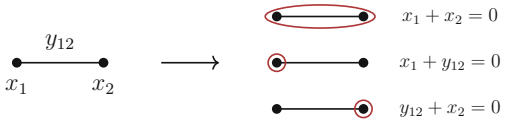
4.2 Wavefunction from Combinatorics

Paolo Benincasa

4.2.1 From Graphs to Polytopes

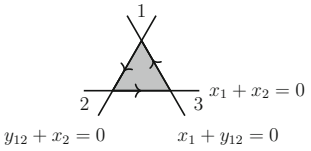
One of the main messages from Sect. 4.1 is the one-to-one correspondence between subgraphs of a given graph $\mathcal{G}$ and singularities of $\psi_{\mathcal{G}}$. In addition, the existence of a recursion relation based on iteratively constructing subgraphs is combinatorial

in nature. So, it becomes legitimate to ask whether there is some mathematical structure which determines it. Let us begin with the simplest example, the two-site tree graph, and let us consider the subgraph (and hence the singularities) associated to it,



(4.2.1)

Note that the loci of the singularities are all homogeneous linear equations in (x_1, y_{12}, x_2). This implies that taking this triple as a local coordinate system in $\mathbb{P}^2$, these three equations determine three lines. Choosing an orientation, the positive half-planes that they identify, i.e., taking $x_1 + x_2 \geqslant 0$, $x_1 + y_{12} \geqslant 0$, $y_{12} + x_2 \geqslant 0$—overlap on a triangle (the shaded area):



(4.2.2)

Interestingly given a triangle $\mathcal{P}_\mathcal{G}$ whose sides are identified by $\mathcal{P}_\mathcal{G} \cap \mathcal{W}^{(\mathcal{G})}$—$\mathcal{W}^{(\mathcal{G})}$ being the normal covector to the line containing one of its sides—there exists a unique (up to normalization) differential form associated to it such that it has only logarithmic singularities and such logarithmic singularities are associated to its boundaries (i.e., sides and vertices):

$$\omega(\mathcal{Y}, \mathcal{P}_\mathcal{G}) = \frac{\langle 123\rangle^2 \langle \mathcal{Y} \mathrm{d}^2 \mathcal{Y}\rangle}{\langle \mathcal{Y}12\rangle\langle \mathcal{Y}23\rangle\langle \mathcal{Y}31\rangle}\,, \tag{4.2.3}$$

where $\mathcal{Y} := (x_1, y_{12}, x_2) \in \mathbb{P}^2$ is a generic point in $\mathbb{P}^2$, $\langle \mathcal{Y} \mathrm{d}^2 \mathcal{Y}\rangle$ is the canonical measure in $\mathbb{P}^2$, $\langle yi(i+1)\rangle \equiv \epsilon_{IJK} y^I Z^J_{(i)} Z^K_{(i+1)}$, $\{Z^I_{(i)}, \quad i = 1, 2, 3\}$ being the set of vertices of the triangle and $\mathcal{W}_\mathcal{I} = \epsilon_{IJK} Z^J_{(i)} Z^K_{(i+1)}$ being the covector that identifies the line passing through the two vertices $Z_{(i)}$ and $Z_{(i+1)}$. The numerator is, in this simple case, just determined by projectivity: the canonical form has to be invariant under GL(1) rescaling $\mathcal{Y} \to \lambda \mathcal{Y}$, $\{Z_{(i)} \to \lambda Z_{(i)}\}$, $\quad i = 1, 2, 3, \lambda \in \mathbb{R}_+$. The coefficient of the canonical form—named canonical function—is nothing but the wavefunction coefficient associated to the two-site tree graph:

$$\omega(\mathcal{Y}, \mathcal{P}_\mathcal{G}) = \Omega(\mathcal{Y}, \mathcal{P}_\mathcal{G}) \langle \mathcal{Y} \mathrm{d}^2 \mathcal{Y}\rangle\,, \tag{4.2.4}$$

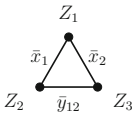
with

$$\Omega(\mathcal{Y}, \mathcal{P}_\mathcal{G}) = \psi_\mathcal{G}(x, y)\,. \tag{4.2.5}$$

The general statement is that given a (reduced) weighted graph $\mathcal{G}$, it is in one-to-one correspondence with a *cosmological polytope*, whose canonical function encodes the wavefunction coefficient associated to $\mathcal{G}$. Going back to the two-site example, we can naturally associate a triangle to it by

(i) considering the local coordinates $\mathcal{Y} \equiv (x_1, y_{12}, x_2) \in \mathbb{P}^2$ constructed out of its weights;
(ii) in this system of local coordinate, considering the canonical basis for $\mathbb{R}^3$, $\mathbf{x}_1 \equiv (1, 0, 0)$, $\mathbf{y}_{12} \equiv (0, 1, 0)$, $\mathbf{x}_2 \equiv (0, 0, 1)$.

Then, the triple $(\mathbf{x}_1, \mathbf{y}_{12}, \mathbf{x}_2)$ defines a triangle whose sides are characterized by $(\mathbf{x}_1, \mathbf{y}_{12}, \mathbf{x}_2)$ as midpoints. The vertices of such a triangle are then given by



$$\begin{aligned} Z_1 &= \mathbf{x}_1 - \mathbf{y}_{12} + \mathbf{x}_2\,, \\ Z_2 &= \mathbf{x}_1 + \mathbf{y}_{12} - \mathbf{x}_2\,, \\ Z_3 &= -\mathbf{x}_1 + \mathbf{y}_{12} + \mathbf{x}_2\,, \end{aligned} \tag{4.2.6}$$

and the associated canonical form is given by

$$\begin{aligned} \omega(\mathcal{Y}, \mathcal{P}_\mathcal{G}) &= \frac{\langle 123\rangle^2 \langle \mathcal{Y} \mathrm{d}^2 \mathcal{Y}\rangle}{\langle \mathcal{Y}12\rangle\langle \mathcal{Y}23\rangle\langle \mathcal{Y}31\rangle} \\ &= \underbrace{\frac{1}{(x_1 + x_2)(x_1 + y_{12})(y_{12} + x_2)}}_{\psi_\mathcal{G}(x,y)} \frac{\mathrm{d}x_1 \wedge \mathrm{d}y_{12} \wedge \mathrm{d}x_2}{\mathrm{Vol}\{\mathrm{GL}(1)\}}\,. \end{aligned} \tag{4.2.7}$$

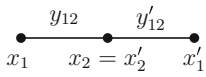
4.2.2 Cosmological Polytopes

It is useful—and perhaps more intuitive—to keep the graphs as our starting point. A first observation is that any weighted graph can be obtained from a collection of two-site graphs by identifying some sites. For example, let us consider a collection of two two-site tree graphs:

(4.2.8)

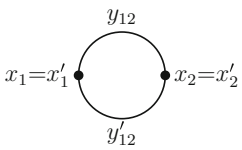
It is possible to obtain from them new graphs from all the inequivalent ways of identifying one or more sites:

1-site identification:



(4.2.9)

2-site identification:

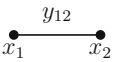


(4.2.10)

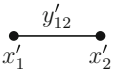
As in each 2-site tree graph there is a triangle associated to it; given two 2-site graphs, it is possible to embed the associated triangles in the same space $\mathbb{P}^5$, where they are still disconnected. The identification of a pair of sites belonging to different 2-site tree graphs, say $x_2 = x'_2$, corresponds geometrically to the identification of the vertices $\mathbf{x}_2 = \mathbf{x}'_2$. In terms of the vertices of the triangles, this implies a linear relation among two pairs of them, i.e., the pair having as a midpoint $\mathbf{x}_2$ and the one having as a midpoint $\mathbf{x}'_2$,

$$Z_3 + Z_1 \sim Z'_3 + Z'_1 \,. \tag{4.2.11}$$

Such a relation identifies a 2-plane—there are as many of such 2-planes as shared sites in a graph. If the 2-site graphs have the following triple associated to them,

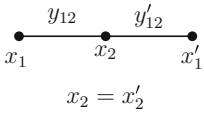


$$\{\mathbf{x_1} - \mathbf{y_{12}} + \mathbf{x_2}, \mathbf{x_1} + \mathbf{y_{12}} - \mathbf{x_2}, -\mathbf{x_1} + \mathbf{y_{12}} + \mathbf{x_2}\}\,, \tag{4.2.12}$$

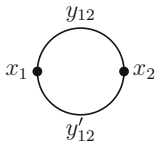


$$\{\mathbf{x'_1} - \mathbf{y'_{12}} + \mathbf{x'_2}, \mathbf{x'_1} + \mathbf{y'_{12}} - \mathbf{x'_2}, -\mathbf{x'_1} + \mathbf{y'_{12}} + \mathbf{x'_2}\}\,, \tag{4.2.13}$$

then the two graph topologies that can be constructed out of these 2-site graphs, have each associated a cosmological polytope with 6 vertices given by



$$\{\mathbf{x_1} - \mathbf{y_{12}} + \mathbf{x_2}, \mathbf{x_1} + \mathbf{y_{12}} - \mathbf{x_2}, -\mathbf{x_1} + \mathbf{y_{12}} + \mathbf{x_2},$$
$$\mathbf{x'_1} - \mathbf{y'_{12}} + \mathbf{x_2}, -\mathbf{x'_1} + \mathbf{y'_{12}} - \mathbf{x_2}, -\mathbf{x'_1} + \mathbf{y'_{12}} + \mathbf{x_2}\} \subset \mathbb{P}^4\,, \tag{4.2.14}$$



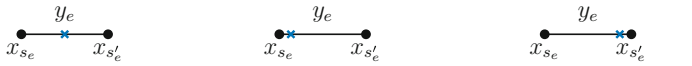
$$\{\mathbf{x_1}-\mathbf{y_{12}}+\mathbf{x_2},\mathbf{x_1}+\mathbf{y_{12}}-\mathbf{x_2},-\mathbf{x_1}+\mathbf{y_{12}}+\mathbf{x_2},$$
$$\mathbf{x_1}-\mathbf{y'_{12}}+\mathbf{x_2},\mathbf{x_1}+\mathbf{y'_{12}}-\mathbf{x_2},-\mathbf{x_1}+\mathbf{y'_{12}}+\mathbf{x_2}\}\subset\mathbb{P}^{\mathbf{3}}\,, \tag{4.2.15}$$

More generally, given an ordinary, connected, graph $\mathcal{G}$ with n_s sites and n_e edges, the associated cosmological polytope lives in $\mathbb{P}^{n_s+n_e-1}$ with local coordinates given by all the weights of the graph $\mathcal{Y}:=(\{x_s\}_{s\in\mathcal{V}},\{y_e\}_{e\in\mathcal{E}}$, and is the convex hull of 3^{n_e} vertices:

$$\{\mathbf{x}_{s_e}-\mathbf{y}_e+\mathbf{x}_{s'_e}\}_{e\in\mathcal{E}}\,. \tag{4.2.16}$$

The correspondence between graphs and cosmological polytopes allows keeping track of the vertex structure of all its boundaries. This can be obtained by introducing a marking that identifies those vertices which *are not* on the facet—a facet is the codimension-1 face of the cosmological polyotope identified by an equation which corresponds to one of the poles of the wavefunction and which are associated to a subgraph.

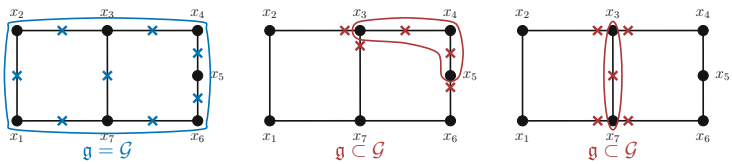
The marking is given by



(4.2.17)

A vertex Z is on a facet identified by a covector $\mathcal{W}$ if $Z^I\mathcal{W}_I=0$. If it is not on the facet, then $Z^I\mathcal{W}_I>0$. This implies that to know which vertex is on a facet, one would have to check whether $Z^I\mathcal{W}_I=0$. A graphical rule in the marking allows us to find the answer immediately.

The facet of $\mathcal{P}_\mathcal{G}$ identified by the subgraph $\mathfrak{g}\subseteq\mathcal{G}$ is given by marking all the internal edges in the middle, and all the edges that depart from this subgraph on the side close to $\mathfrak{g}$:



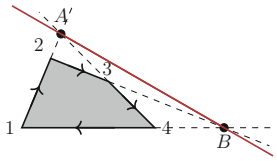
Due to the one-to-one mapping between subgraphs and facets, the latter can be classified or listed using the former. The canonical functions of a cosmological

polytope can generally be written as follows:

$$\omega(\mathcal{Y}, \mathcal{P}_{\mathcal{G}}) = \frac{\mathfrak{n}_\delta(\mathcal{Y})}{\prod_{\mathfrak{g}\subseteq\mathcal{G}} q_{\mathfrak{g}}(\mathcal{Y})} \langle \mathcal{Y} \mathrm{d}^{n_s+n_e} \mathcal{Y} \rangle \,, \tag{4.2.18}$$

where $q_{\mathfrak{g}} \equiv \mathcal{Y} \cdot \mathcal{W}^{(g)}$ and $\mathfrak{n}_\delta(\mathcal{Y})$ is a polynomial of degree δ, with δ fixed by GL(1) invariance to be $\delta = \tilde{\nu} - n_s - n_e$ ($\tilde{\nu} \equiv$ number of facets). It turns out that also the polynomial $\mathfrak{n}_\delta(\mathcal{Y})$ has a geometrical interpretation: it is the locus of the intersections of the hyperplanes containing the facet of $\mathcal{P}_{\mathcal{G}}$ outside $\mathcal{P}_{\mathcal{G}}$—in the mathematical literature $\mathfrak{n}_\delta(\mathcal{Y}) = 0$ defines the adjoint surface.

A toy example can be given by a square in $\mathbb{P}^2$:



(4.2.19)

The square is the convex hull of the vertices 1, 2, 3, 4 and is characterized by the following canonical form

$$\omega(\mathcal{Y}, \mathcal{P}) = \frac{\mathfrak{n}_\delta(\mathcal{Y}) \langle \mathcal{Y} \mathrm{d}^2 \mathcal{Y} \rangle}{\langle \mathcal{Y}12 \rangle \langle \mathcal{Y}23 \rangle \langle \mathcal{Y}34 \rangle \langle \mathcal{Y}41 \rangle} \,. \tag{4.2.20}$$

Because of GL(1) invariance, the numerator ought to be a polynomial of degree $\delta = 1$. If higher codimension faces are defined as intersections of lower-codimension ones, and hence manifest themselves in terms of non-vanishing multiple residue of the canonical form—there are further intersections which do not define a higher codimension face (like the points A and B for the quadrilateral above): the sequential residue of the canonical form has to vanish. This implies that these intersections define subspaces of the adjoint. The numerator can be computed knowing these intersections. In the case of the quadrilateral above,

$$Z^I_A = \epsilon^{IJK} \mathcal{W}^{(12)}_J \mathcal{W}^{(34)}_K \,, \qquad Z^I_B = \epsilon^{IJK} \mathcal{W}^{(23)}_J \mathcal{W}^{(41)}_K \,, \tag{4.2.21}$$

where $\mathcal{W}^{(ij)}_J$ identifies the line through the vertices $Z^I_{(i)}$ and $Z^I_{(j)}$ and

$$\mathfrak{n}_1(\mathcal{Y}) = \mathcal{Y}^I \epsilon_{IJK} Z^J_A Z^K_B \,. \tag{4.2.22}$$

The knowledge of these intersections determine also compatibility conditions for the facets. While for the simple example of the square is trivial, determining all the compatibility conditions of all codimension is not an easy task. In the case of the cosmological polytopes, the correspondence between graphs and polytopes,

together with the markings, allow unveiling all this structure. From a physics perspective, these conditions imply which sequential discontinuities are non-zero. In codimension-two, they allowed to determine the analog for the wavefunction of the Steinmann relations [7]. The higher-codimension one allows for novel ways of representing the wavefunction without introducing spurious poles [8]. In order to fix the ideas, let us go back to the example of the quadrilateral. Rather than fixing the numerator, one could determine the canonical form from a triangulation,

$$\omega(\mathcal{Y},\mathcal{P}) = \omega(\mathcal{Y},\mathcal{P}_{124}) + \omega(\mathcal{Y},\mathcal{P}_{423}) = \omega(\mathcal{Y},\mathcal{P}_{123}) + \omega(\mathcal{Y},\mathcal{P}_{341})\,. \tag{4.2.23}$$

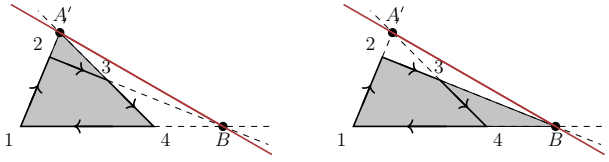
(4.2.24)

Note that in these triangulations, one makes use of a new codimension-1 boundary (the segments 24 and 13).

From the perspective of the canonical forms as well as the codimension-1 boundary correspond to a singularity, the presence of (24) and (13) introduce a spurious singularity in each of the two representations.

In order to avoid the introduction of spurious boundaries/singularities, one should use only triangulations via the actual facets of the polytope of interest. In the case of the quadrilateral,

$$\omega(\mathcal{Y},\mathcal{P}) = \omega(\mathcal{Y},\mathcal{P}_{1A4}) + \omega(\mathcal{Y},\mathcal{P}_{3A2}) = \omega(\mathcal{Y},\mathcal{P}_{12B}) + \omega(\mathcal{Y},\mathcal{P}_{B34})\,, \tag{4.2.25}$$



(4.2.26)

i.e., one of the triangulations via one of the subspaces of the endpoint. In the case of the cosmological polytope, both the knowledge of the subspaces and of the compatibility conditions, is made possible by the markings, allowing to classify all these representations. The structure can be understood via the simple example above,

$$\omega(\mathcal{Y},\mathcal{P}) = \omega(\mathcal{Y},\mathcal{P}_{1A4}) + \omega(\mathcal{Y},\mathcal{P}_{3A2}) = \frac{1}{q_{12}q_{34}}\left[\frac{1}{q_{23}} + \frac{1}{q_{41}}\right]. \tag{4.2.27}$$

The prefactor $\frac{1}{q_{12}q_{34}}$ determines the subspace through which the triangulation is performed, and one then sums over all the compatible boundaries, as represented by the terms in the square bracket.

4.3 Asymptotic Structure of Cosmological Integrals

Francisco Vazão

The goal of this lecture is to extract IR divergences for any cosmological integral in FRW cosmology. We will describe in detail the integrals that arise in cosmology before providing an example of how to compute them. We will be able to show how to do it for a general tree-level integral, and describe an example of a one-loop integral before concluding.

4.3.1 Infrared Effects in FRW

As a simple example of how infrared effects in FRW come about, we take a massless scalar with potential $V(\phi) = \lambda\phi^4$ in four spacetime dimensions in de Sitter spacetime, i.e., in dS_{1+3}. The equal-time two-point function behaves roughly like

$$\langle\phi_{\mathbf{k}}(\eta)\phi_{\mathbf{k}'}(\eta)\rangle^{\text{1-loop}} \sim \frac{H^2}{k^3}\left[1+\lambda\log\left(\frac{k}{Ha(\eta)}\right)\log\left(kHa(\eta)L^2\right)\right], \tag{4.3.1}$$

where L is the infrared cutoff, $a(\eta) \sim \frac{1}{\eta H}$ is the warp factor in de Sitter, k is the magnitude of momentum of the fields and H is the Hubble constant. There are two sources of infrared divergences in this expression. The first one is called a secular logarithm: it is an effect of putting QFT in an expanding spacetime, causing an accumulation of long wavelength modes in the future, as $\eta \to 0^-$. In this limit, the second term also goes to infinity. In addition, we have an IR divergence that comes from integrating the loop momentum over low energies, just like in flat space, which will occur upon an integration over k.

In examples like this one where infrared divergences are present, one of three things might happen: one cannot do perturbation theory, infrared divergences could cancel between different contributions, or one can resum an infinite set of diagrams to produce a sensible answer. So, some of the questions we want to answer here are:

(i) Which conditions do the IR divergences have to satisfy in order to admit a resummation?
(ii) What are the completely IR finite observables in a general FRW cosmology?

In Starobinsky's stochastic inflation, the Focker-Planck equation tells us that the probability distribution function reaches equilibrium, and it has been shown that the leading logarithm matches the solution given by Starobinsky [9, 10]. There are still some questions, however, e.g., what happens in general FRW cosmology and how to resum subleading logarithms. The goal of this lecture is to understand

these questions in a more general setting, without restricting ourselves to a specific cosmology.

Audience Question 4.3.1. There has already been a lot of work on this in de Sitter, so how does this talk relate to prevoius work?
Answer: The physical picture is that we are not starting in the free vacuum. There have already been many papers on this in de Sitter, showing that stochastic inflation resums logarithms and the system has a fixed point. We need to choose the correct state in the past, which we take to be the Bunch–Davies vacuum. How do we input the state? It is a fixed point (dynamical), so there is no simple pure state to input.

4.3.2 Cosmological Integrals

We take the warp factor of FRW to be $a(\eta) = \frac{1}{\eta^\gamma}$, where $\gamma = 1$ corresponds to de Sitter but here we take γ to be generic. Then, an integral corresponding to a graph $\mathcal{G}$ in perturbation theory takes the generic form

$$\mathcal{I}_{\mathcal{G}}[\alpha,\ \beta;\ \mathcal{X}] = \int\limits_{\mathbb{R}_+^2} \prod_{s\in\mathcal{V}} \left[\frac{\mathrm{d}x_s}{x_s}\, x_s^{\alpha_s}\right] \int_\Gamma \prod_{e\in\mathcal{E}^{(L)}} \left[\frac{\mathrm{d}y_e}{y_e}\, y_e^{\beta_e}\right] \times \frac{\mu_d(y_e;\mathcal{X})^{\frac{d-L-n_s}{2}}\, \mathfrak{n}_\delta(z,\ \mathcal{X})}{\prod\limits_{\mathfrak{g}\subseteq\mathcal{G}} \left[q_{\mathfrak{g}}(z,\ \mathcal{X})\right]^{\tau_{\mathfrak{g}}}}\,. \tag{4.3.2}$$

Here, μ_d is the loop integration measure in d spatial dimensions, which is polynomial in the loop integration variables y_e and also depends on the external kinematics $\mathcal{X}$. The parameters α_s depend on cosmology and coupling as follows:

$$\alpha = \gamma\Big[2 - \frac{(\kappa-1)(d-1)}{2}\Big] + \sum_s \ell_s\,, \tag{4.3.3}$$

where s are the sites of $\mathcal{G}$ and $\mathcal{V}$ is the set of all sites in $\mathcal{G}$. The β_e depend on which wavefunction we are computing. n_s is the total number of sites in $\mathcal{G}$, and $\mathfrak{g}$ is a subgraph of $\mathcal{G}$. The numerator $\mathfrak{n}_\delta$ is a polynomial of degree δ. We have labeled with $z \equiv (x, y)$ the vector that contains the integration variables $x \equiv (x_s)_{s\in\mathcal{V}}$ and $y \equiv (y_e)_{e\in\mathcal{E}^{(L)}}$. Note that $\mathcal{E}^{(L)}$ is the set of edges of $\mathcal{G}$ that have at least one loop momentum running through it. Γ is an integration contour that is given by the condition $\mu_d \geqslant 0$.

These types of integrals appeared in the previous subsections of this chapter, and our goal in this section is to analyze their infrared divergences.

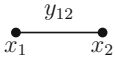
Example 4.3.1. The integration measure for the bubble integral is

$$\mu_{\rm bub} \sim \begin{vmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & y_a^2 & y_b^2 \\ 1 & y_a^2 & 0 & p^2 \\ 1 & y_b^2 & p^2 & 0 \end{vmatrix} = \mathrm{Vol}^2 \left(\begin{array}{c} y_b \quad\quad y_a \\ \triangle \\ p \end{array} \right) .$$

(4.3.4)

4.3.3 The Two-Site Chain

In this section, we go through an example of how to derive the IR divergences of the two-site chain,



(4.3.5)

The integral is given by

$$\mathcal{I}_2^{(0)}[\alpha] \equiv \int_0^\infty \frac{\mathrm{d}x_1}{x_1}\, x_1^\alpha \int_0^\infty \frac{\mathrm{d}x_2}{x_2}\, x_2^\alpha \frac{1}{q_{\mathfrak{g}} q_{\mathfrak{g}_1} q_{\mathfrak{g}_2}} . \tag{4.3.6}$$

with

$$q_{\mathfrak{g}} = x_1 + x_2 + X_1 + X_2 , \tag{4.3.7}$$

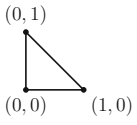
$$q_{\mathfrak{g}_2} = x_1 + y + X_1 , \tag{4.3.8}$$

$$q_{\mathfrak{g}_2} = x_2 + y + X_2 . \tag{4.3.9}$$

We use the *Newton Polytope* to analyze the different ways in which the denominator factors can go to zero. Recall that X_1 and X_2 are external energies, but x_1 and x_2 are integrated over. We therefore form the Newton polytope of the polynomial

$$q_{\mathfrak{g}} = x_1^1 x_2^0 + x_1^0 x_2^1 + x_1^0 x_2^0 (X_1 + X_2) , \tag{4.3.10}$$

as the polytope with vertices at the powers of each monomial that appears:



(4.3.11)

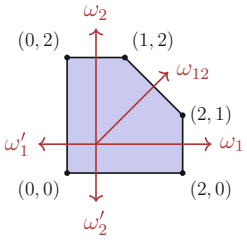
To find the Newton polytope of a product of polynomials, such as the product $q_{\mathfrak{g}} q_{\mathfrak{g}_1} q_{\mathfrak{g}_2}$ appearing in the denominator of (4.3.6), we can either multiply out all the factors or simply take the Minkowski sum of the polytopes that we found for each polynomial. That is, we sum their vertices. For the bubble example, we get,

$$N(q_{\mathfrak{g}} q_{\mathfrak{g}_1}) = N(q_{\mathfrak{g}}) \oplus N(q_{\mathfrak{g}_1}) = \tag{4.3.12}$$

and

$$N(q_{\mathfrak{g}} q_{\mathfrak{g}_1} q_{\mathfrak{g}_2}) = N(q_{\mathfrak{g}} \oplus N(q_{\mathfrak{g}_1}) \oplus N(q_{\mathfrak{g}_2}) = \tag{4.3.13}$$

The integral $\mathcal{I}$ will converge when the point (α, α) is inside the Newton polytope. Next, we find the vectors normal to the facets of this Newton polytope, which we label with ω_i:



(4.3.14)

This pentagon is a nestohedron with five different directions that could possibly be divergent, each direction given a ω.

We can now use the Newton polytope to understand the divergences of the integral (4.3.6). We embed the Newton polytope N in $\mathbb{P}^2$ by putting $Z_i = \binom{1}{\rho_i}$. We also introduce the covectors

$$\mathcal{W}_I^{ab} = \epsilon_{I J_1 J_2} Z_a^{J_1} Z_b^{J_2} = \begin{pmatrix} \lambda^{ab} \\ \boldsymbol{\omega}^{ab} \end{pmatrix}, \tag{4.3.15}$$

with $\lambda^{ab} = (\alpha, \alpha) \cdot \boldsymbol{\omega} - \max_\rho(\boldsymbol{\rho} \cdot \boldsymbol{\omega})$. The purpose of adding λ^{ab} to the covectors is that $\lambda \geqslant 0$ indicates that the integral $\mathcal{I}$ will diverge. The covectors for the two-site example are explicitly given by,

$$\mathcal{W}'^{(1)} = \begin{pmatrix} -\alpha \\ -1 \\ 0 \end{pmatrix}, \quad \mathcal{W}'^{(2)} = \begin{pmatrix} -\alpha \\ 0 \\ -1 \end{pmatrix}, \tag{4.3.16}$$

$$\mathcal{W}^{(1)} = \begin{pmatrix} \alpha-2 \\ 1 \\ 0 \end{pmatrix}, \quad \mathcal{W}^{(2)} = \begin{pmatrix} \alpha-2 \\ 0 \\ 1 \end{pmatrix}, \quad \mathcal{W}^{(12)} = \begin{pmatrix} 2\alpha-3 \\ 1 \\ 1 \end{pmatrix}. \tag{4.3.17}$$

We can therefore read off the first entries that the integral $\mathcal{I}$ will be divergent if $2\alpha - 3 \geqslant 0$, $-\alpha \geqslant 0$ or $\alpha - 2 \geqslant 0$. An equal sign denotes a logarithmic divergence, while a strict inequality signals a power-law divergence.

Let us take $2\alpha - 3 \to 0$, which corresponds to a logarithmic divergence of (4.3.6). From the covectors in (4.3.17), we see that the only divergent direction is that of $\mathcal{W}^{(12)}$, since the first entry of the other covectors is always negative for this choice. In this case, the integral receives contributions from two divergent regions, which share the covector $\mathcal{W}^{(12)}$ but differ by including $\mathcal{W}^{(1)}$ or $\mathcal{W}^{(2)}$.

Next, we use sector decomposition to compute the infrared divergences of the integral. In the first sector, labeled by $\mathcal{W}^{(12)}$ and $\mathcal{W}^{(1)}$ we change integration variables to

$$x_j \equiv \zeta_{12}^{-\mathfrak{e}_j \cdot \omega_{12}} \zeta_1^{-\mathfrak{e}_j \cdot \omega_1}, \qquad \text{for } j = 1, 2\,, \tag{4.3.18}$$

where $\mathfrak{e}_j$ is a unit vector in the direction of ω_j. The integral from (4.3.6) can then be rewritten as

$$\begin{aligned} \mathcal{I}_{\Delta_{\text{IR}}^{(1)}} \;=\; & \int_0^1 \frac{\mathrm{d}\zeta_1}{\zeta_1} \frac{\mathrm{d}\zeta_{12}}{\zeta_{12}} \\ & \times \frac{\zeta_1^{-\lambda^{(1)}} \zeta_{12}^{-\lambda^{(12)}}}{[1+\zeta_1+(X_1+X_2)\zeta_{12}\zeta_1]\,[1+(y+X_1)\zeta_{12}\zeta_1]\,[1+(y+X_2)\zeta_{12}]}\,, \end{aligned} \tag{4.3.19}$$

with $\lambda^{(12)} = 2\alpha - 3$ and $\lambda^{(1)} = \alpha - 2$. The divergence when $\lambda^{(12)} \to 0$ comes from the integration region where $\zeta_{12} \to 0$. Thus, when computing the leading divergence of $\mathcal{I}_{\Delta_{\text{IR}}^{(1)}}$, the change of variables in (4.3.18) allows us to drop subleading terms in ζ_{12}. The divergent part is therefore simply given by

$$\mathcal{I}^{\text{div}}_{\Delta_{\text{IR}}^{(1)}} \;=\; -\frac{1}{\lambda^{(12)}} \int_0^1 \frac{\mathrm{d}\zeta_1}{\zeta_1} \frac{\zeta_1^{1/2}}{(1+\zeta_1)}\,, \tag{4.3.20}$$

where we have evaluated the coefficient of the pole in $\lambda^{(12)}$ at $\lambda^{(12)} = 0$, resulting in $\lambda^{(1)}|_{\lambda^{(12)}=0} = -1/2$. The remaining integral is straightforward to evaluate, and we obtain

$$\mathcal{I}^{\text{div}}_{\Delta_{\text{IR}}^{(1)}} \;=\; -\frac{\pi}{2\,\lambda^{(12)}}\,. \tag{4.3.21}$$

The divergent contribution can be rewritten by restoring projective invariance as

$$\mathcal{I}^{\text{div}}_{\Delta_{\text{IR}}^{(1)}} \;=\; -\frac{1}{2\lambda^{(12)}} \int_{\mathbb{R}_2^+} \frac{\mathrm{d}x_1}{x_1} \frac{\mathrm{d}x_2}{x_2}\, x_1^{\alpha-1}\, x_2^{\alpha-1} \frac{1}{\text{Vol}\{\text{GL}(1)\}} \frac{1}{x_1 + x_2}\,, \tag{4.3.22}$$

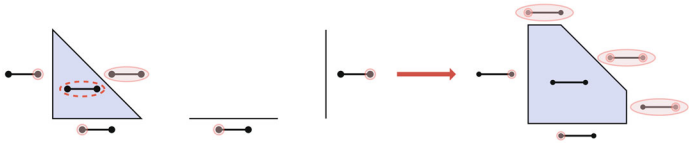
with $\alpha = 3/2$, which we can interpret as an integral corresponding to a graph with a single site of weight $x_1 + x_2$.

When adding the two divergent sectors, we can therefore write the total divergent contribution as

$$\mathcal{I}^{\text{div}}_{\Delta^{(1)}} + \mathcal{I}^{\text{div}}_{\Delta^{(2)}} = -\frac{1}{\lambda^{(12)}} \int_{\mathbb{R}^+_2} \frac{\mathrm{d}x_1}{x_1} \frac{\mathrm{d}x_2}{x_2} \frac{x_1^\alpha\, x_2^\alpha}{\text{Vol}\{\text{GL}(1)\}} \Omega(x_1, x_2, X_1 = 0, X_2 = 0)\,, \tag{4.3.23}$$

with $\alpha = 3/2$.

Note that we can bypass the steps above if we instead form a graph associahedron for the tree graphs. For example, the associahedron for the two-site chain is:



(4.3.24)

Comparing this diagram with (4.3.14) shows that we have recovered the Newton Polytope for the graph.

4.3.4 General Tree Graphs

We would now like to generalize the construction in the previous section to any tree graph. A generic tree graph is given by the expression,

$$\mathcal{I}^{\text{tree}}_{\mathcal{G}}[\alpha,\, \tau,\, \mathcal{X}] = \int_0^{+\infty} \prod_{s\in\mathcal{V}} \left[\frac{\mathrm{d}x_s}{x_s}\, x_s^{\alpha_s}\right] \frac{\mathfrak{n}_\delta(x,\, \mathcal{X})}{\prod_{\mathfrak{g}\subseteq\mathcal{G}} \left[q_{\mathfrak{g}}(x,\, \mathcal{X})\right]^{\tau_{\mathfrak{g}}}}\,, \tag{4.3.25}$$

where we have used the same definitions as around (4.3.2).

Note that the Newton polytope corresponding to each denominator factor $\mathcal{N}(q_{\mathfrak{g}})$ is a simplex. Thus, we can form their Minkowski sum as

$$\mathcal{N}_{\mathcal{G}} \equiv \bigoplus_{\mathfrak{g}\subseteq\mathcal{G}} \text{Re}\{\tau_{\mathfrak{g}}\}\Sigma[\mathfrak{g}]\,. \tag{4.3.26}$$

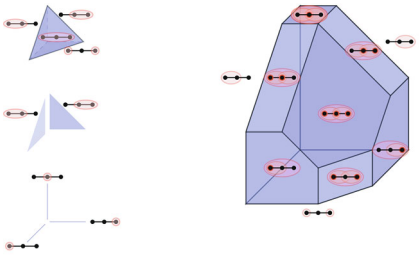
Schematically, it can be realized by taking the highest-dimensional simplex and truncating it with the lower-dimensional ones. Then, for all the facets that do not identify with the coordinate planes, one can associate a nested tubing consisting of all tubings which contribute in the normal direction to the facet (as illustrated in (4.3.28)). With the Newton polytope for the graph at our disposal, we can write

down the co-vectors, which are given by

$$W^{(j_1 \dots j_{n_s^{(\mathfrak{g})}})} = \begin{pmatrix} \sum_s \alpha_s - (\#\text{ number of nested tubings}) \\ \mathfrak{e}_{j_1 \dots j_{n_s^{(\mathfrak{g})}}} \end{pmatrix}, \tag{4.3.27}$$

where $\mathfrak{e}_{j_1 \dots j_{n_s^{(\mathfrak{g})}}}$ is a vector with 1 in the entries $j_1 \dots j_{n_s^{(\mathfrak{g})}}$ and zero in all other entries. Thus, we have found a diagrammatic starting point for using Sector decomposition.

Example 4.3.2. The Newton polytope of the three-site chain is formed as follows:



(4.3.28)

4.3.5 Loop Diagrams

In this section, we look at an example of a one loop two-site diagram, which we refer to as bubble diagram. Its expression is

$$\mathcal{I}_{\mathcal{G}}^{\text{bubble}} = \prod_{s \in \mathcal{V}} \left[\frac{\mathrm{d}x_s}{x_s} x_s^{\alpha} \right] \int_{\Gamma} \prod_{e \in \mathcal{E}} \left[\frac{\mathrm{d}y_e}{y_e} y_e^{\beta} \right] \frac{\left[-\mu(y^2, P^2) \right]^{\frac{d-3}{2}}}{P^{d-2}} \Omega(x, y; \mathcal{X}) , \tag{4.3.29}$$

where we use the same definitions as around (4.3.2) and the integration measure was given in (4.3.4). The denominator factors are explicitly given by

$$\Omega(x, y; \mathcal{X}) = \frac{\mathfrak{n}_{\delta}}{q_{\mathcal{G}}^{\tau_{\mathcal{G}}} q_{\mathfrak{g}_a}^{\tau_{\mathfrak{g}_a}} q_{\mathfrak{g}_b}^{\tau_{\mathfrak{g}_b}} q_{\mathfrak{g}_1}^{\tau_{\mathfrak{g}_1}} q_{\mathfrak{g}_2}^{\tau_{\mathfrak{g}_1}}} , \tag{4.3.30}$$

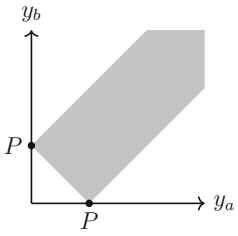
with

$$\begin{aligned} q_{\mathfrak{g}_a} &= x_1 + x_2 + 2y_a + X_1 + X_2, \quad q_{\mathfrak{g}_b} = x_1 + x_2 + 2y_b + X_1 + X_2, \\ q_{\mathfrak{g}_1} &= x_1 + y_a + y_b + X_1, \quad q_{\mathfrak{g}_2} = x_2 + y_a + y_b + X_2 \quad q_{\mathcal{G}} = x_1 + x_2 + X_1 + X_2 . \end{aligned} \tag{4.3.31}$$

The integration contour is

$$\Gamma = \{y^2 : \mu(y^2) \geqslant 0\}\,, \tag{4.3.32}$$

which from (4.3.4) gives the following region of integration:



(4.3.33)

It is clear from the integration region that whenever we set one of the edge variables to zero the other is set to P. This can also be seen from the perspective of the integration measure, which is the squared volume of a simplex. For the bubble we have a triangle, by setting one edge to zero the other two edges collapse into each other. For higher site graphs a similar behavior is seen, setting one edge to zero leads all others to some finite value which is a function of the external kinematics. On the other end, if we take the value of one edge variable to infinity, then all others are forced to go to infinity. This essentially fixes all the singularities of the loop integration for our bubble integral, they are $\{y_a \to 0, y_b \to P\}$, $\{y_a \to P, y_b \to 0\}$ and $\{y_a \to \infty, y_b \to \infty\}$. For each of these points we can analyze their neighbor region, with a suitable change of variables, and perform sector decomposition (together with the site integration) to extract the divergences.

References

1. D. Anninos, De sitter musings. Int. J. Mod. Phys. A **27**, 1230013 (2012). [1205.3855]
2. P. Benincasa, Amplitudes meet cosmology: a (scalar) primer (2022). 2203.15330
3. P. Benincasa, Amplitudes meet cosmology, in *Recorded Lectures from Workshop on Scattering Amplitudes and Cosmology at ICTP* (2014)
4. P. Benincasa, F. Vazão, The asymptotic structure of cosmological integrals (2022). 2402.06558
5. M. Baumgart, R. Sundrum, Manifestly causal in-in perturbation theory about the interacting vacuum. J. High Energy Phys. **03**, 080 (2021). [2010.10785]
6. S. Albayrak, P. Benincasa , C. Duaso Pueyo, Perturbative unitarity and the wavefunction of the Universe. SciPost Phys. **16**, 157 (2024). [2305.19686]
7. P. Benincasa, A.J. McLeod, C. Vergu, Steinmann relations and the wavefunction of the universe. Phys. Rev. D **102**, 125004 (2020). [2009.03047]
8. P. Benincasa, W.J.T. Bobadilla, Physical representations for scattering amplitudes and the wavefunction of the universe. SciPost Phys. **12**, 192 (2022). [2112.09028]
9. A.A. Starobinsky, Stochastic de sitter (inflationary) stage in the early universe. Lect. Notes Phys. **246**, 107 (1986)
10. A.A. Starobinsky, J. Yokoyama, Equilibrium state of a selfinteracting scalar field in the De Sitter background. Phys. Rev. D **50**, 6357 (1994). [astro-ph/9407016]

Asymptotic Observables: The Analytic S-Matrix Revisited

5

Simon Caron-Huot and Mathieu Giroux

Abstract

These lectures introduce the notion of asymptotic observables, which are classes of measurable quantities predicted by quantum field theory. In gapped theories with trivial infrared dynamics, these include scattering amplitudes, expectation values of operators approaching infinity, inclusive cross sections, etc. We argue for a unified picture in which various asymptotic observables are related by analytic continuation embodying new versions of crossing symmetry. As an application, we discuss the exponentiation of infrared divergences for inclusive observables in Quantum Electrodynamics using time-folded contours. Finally, we outline prospects for a systematic study of analyticity properties of asymptotic observables using the Fourier-Bros-Iagolnitzer transform.

5.1 Asymptotic Observables and Crossing

Simon Caron-Huot

In this first lecture, we introduce a minimal set of axioms that allow us to enumerate asymptotic observables of a given multiplicity. We will then explain why these are, remarkably, interconnected through crossing symmetry. The main references on which these lectures are based are [2, 3].

S. Caron-Huot (✉) · M. Giroux
Department of Physics, McGill University, Montréal, QC, Canada
e-mail: simon.caron_huot@mcgill.ca; mathieu.giroux2@mail.mcgill.ca

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

N. Arkani-Hamed et al. (eds.), *Records from the S-Matrix Marathon*,
Lecture Notes in Physics 1041, https://doi.org/10.1007/978-3-031-90352-6_5

5.1.1 Summary of Conventions

First, we spell out here the conventions assumed in this chapter, which follow [3]. Unless specified otherwise, we work in D spacetime dimensions and take the momentum components of p^μ to be

$$p^\mu = (p^0, \underbrace{p^1, \ldots, p^{\mathrm{D}-1}}_{=\mathbf{p}}) . \tag{5.1.1}$$

We work in mostly-plus signature such that $p^2 = -(p^0)^2 + (p^1)^2 + \ldots + (p^{\mathrm{D}-1})^2$ and all-outgoing conventions, in which $p_i^0 < 0$ for the incoming particles and $p_i^0 > 0$ for the outgoing ones. When using lightcone coordinates, we define

$$p^\pm = p^0 \pm p^{\mathrm{D}-1} , \quad \mathbf{p}^\perp = (p^1, \ldots, p^{\mathrm{D}-2}) , \tag{5.1.2}$$

such that $p^2 = -p^+ p^- + (\mathbf{p}^\perp)^2$. Furthermore, we use the notation $p^\mu_{ij\cdots k} = p_i^\mu + p_j^\mu + \ldots + p_k^\mu$, and $s_{ij\cdots k} = -p^2_{ij\cdots k}$ for Mandelstam invariants.

In what follows, the *connected interacting part* T of the S-matrix is obtained in the usual way:

$$S = \mathbb{1} + iT + \ldots , \tag{5.1.3}$$

where "…" include disconnected products of T's and are only present for multiplicities greater than $2 \leftarrow 2$. The interacting part of the *scattering amplitude*, $\mathcal{M}$, is defined as the matrix element of T with the overall momentum-conserving delta function stripped out:

$$\mathcal{M}_{f\leftarrow i} = \langle f|T|i\rangle (2\pi)^{\mathrm{D}} \delta^{\mathrm{D}}\Big(\sum_i p_i^\mu\Big) . \tag{5.1.4}$$

When drawing diagrams, we will follow the operator ordering in bra-ket notation (see, e.g., Eqs. (5.1.23), (5.1.25a), (5.1.25b) below). This means (thanks Dirac!) that time flows toward the *left* in scattering amplitudes S or $\mathcal{M}$, and toward the right in conjugated amplitudes $S^\dagger$ or $\mathcal{M}^\dagger$ factors.

Finally, "in" scattering states are denoted without explicit labels: $\langle \ldots | = {}_{\text{in}}\langle \ldots |$ and $|\ldots\rangle = |\ldots\rangle_{\text{in}}$. In contrast, the "out" states will remain labeled: ${}_{\text{out}}\langle \ldots |$ and $|\ldots\rangle_{\text{out}}$.

5.1.2 Asymptotic Observables

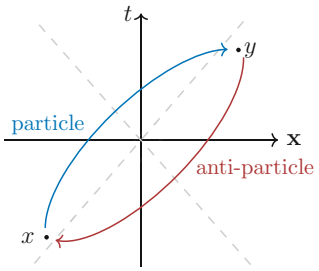
The discussion that follows is motivated by the observation that, in local quantum field theory, causality admits a precise statement in terms of *microcausality*. It

asserts that operators corresponding to observables at spacelike separated points must commute:

$$[O(x), O(y)] = 0 \qquad \text{(spacelike)}\,, \tag{5.1.5}$$

ensuring that it is impossible to send signals faster than light.

Vacuum matrix elements of such relations, $\langle 0|\,[\Phi(y), \Phi^\dagger(x)]\,|0\rangle = 0$ are well known to imply that any particle must have an antiparticle of the same mass and spin. That is, the two terms in the commutator can only cancel out if the amplitude for a Φ particle to move from x to y (which is nonvanishing spacelike) coincides with that of a corresponding antiparticle to go from y to x:



(5.1.6)

The question we wish to explore below is the following: What are the implications of (5.1.5) inside a general scattering state? As we shall see, this will relate the amplitudes of particles and those of antiparticles: *crossing symmetry*, but the answer will also reveal new and non-trivial relations among *out-of-time-order* (OTO) observables. The OTO nature of crossing can already be anticipated, since commutators alter the ordering of the operators.

In time-ordered perturbation theory, the relation (5.1.5) (and the analogous anti-commutation relation for fermions) ensures Lorentz invariance of T-products [4]. Here we will directly use (5.1.5), which seems to buy us more mileage at the end of the day.

Next, let us outline the assumptions used to make this discussion more precise and to derive the results below.

5.1.2.1 S-matrix Axioms

The idealized setup that epitomizes the idea of the S-matrix is a quantum field theory (QFT) with trivial infrared dynamics (e.g., pion scattering in D = 4, QED and QCD in dimensional regularization, etc.). In this setup, it makes sense to *assume* that any finite energy excitation decays at late times into a finite set of stable particles. These particles then separate from each other and effectively become free as they cease to overlap. This assumption and its consequences are encoded in the following set of axioms [3] (similar sets of axioms have a long history [5]):

(i) The algebra of asymptotic measurements in the *far past* is generated by *creation* and *annihilation operators* of stable particles, satisfying the canonical relation:

$$[a_i, a_j^\dagger] = \delta_{i,j}\, 2p_i^0\, (2\pi)^{\mathrm{D}-1}\delta^{\mathrm{D}-1}(\mathbf{p}_i - \mathbf{p}_j) \;\text{ and }\; [a_i, a_j] = [a_i^\dagger, a_j^\dagger] = 0\,. \tag{5.1.7}$$

Here, p_i^0 is the (positive) energy of the ith particle and $\delta_{i,j}$ denotes a Kronecker delta in the flavor and spin indices.

(i') There is an equivalent algebra of "out" measurements in the *far future*, which we denote by b and $b^\dagger$'s. Translated to textbook conventions, the above creation and annihilation operators correspond to

$$a_i \equiv a_i^{\text{in}} \quad \text{and} \quad b_i \equiv a_i^{\text{out}}\,. \tag{5.1.8}$$

(ii) These operators act on equivalent Hilbert spaces and are related by a unitary *evolution operator* S:

$$b_i = S^\dagger a_i S, \qquad b_i^\dagger = S^\dagger a_i^\dagger S\,. \tag{5.1.9}$$

It follows from (5.1.9) that $[a, b]$ can be a complicated object. This is expected because measurements in the past affect the future in a complicated way.

(iii) There exists a *time-invariant vacuum*, $|0\rangle$, that does not contain particles:

$$a_i|0\rangle = b_i|0\rangle = 0, \qquad S|0\rangle = |0\rangle\,. \tag{5.1.10}$$

(iv) One-particle states evolve trivially, which means that they are *stable*

$$b_i^\dagger|0\rangle = S^\dagger a_i^\dagger|0\rangle = a_i^\dagger|0\rangle\,. \tag{5.1.11}$$

This also implies $Sa_i^\dagger|0\rangle = a_i^\dagger|0\rangle$ from the unitarity condition $S^\dagger S = \mathbb{1}$.

Given these axioms and the algebra of measurements, we can now ask: what can be measured asymptotically? We organize the discussion by multiplicity, starting at four-point.

5.1.2.2 Enumerating Observables: *Four-Point*

The possible four-point measurements are of the form

$$\langle 0|\ldots|0\rangle \quad \text{where "}\ldots\text{" is any length-four string of } a, a^\dagger, b, b^\dagger\,. \tag{5.1.12}$$

Naturally, numerous of these combinations result in trivial observables. This is because $a\,|0\rangle = b\,|0\rangle = 0$, and both $a_i a_j^\dagger\,|0\rangle$ and $b_i b_j^\dagger\,|0\rangle$ are proportional to a

delta function, which vanishes for non-forward momenta (and in any rate is a c-number times $|0\rangle$ according to (5.1.7): an uninteresting state!). Therefore, only two non-trivial states can be generated by acting with two asymptotic operators on the vacuum:

$$a_i^\dagger a_j^\dagger|0\rangle \equiv |ij\rangle \quad \text{and} \quad b_i^\dagger a_j^\dagger|0\rangle = b_i^\dagger b_j^\dagger|0\rangle \equiv |ij\rangle_{\text{out}} \,. \tag{5.1.13}$$

These are the *only* two options! Multiplying by the analogous classification on the left, we conclude that the possible observables are therefore the conventional (causal) scattering amplitude

$$\langle 0|b_4 b_3 a_2^\dagger a_1^\dagger|0\rangle = \langle 0|a_4 a_3 S a_2^\dagger a_1^\dagger|0\rangle = \langle 34|\, S\, |12\rangle = {}_{\text{out}}\langle 34|12|34|12\rangle \,, \tag{5.1.14}$$

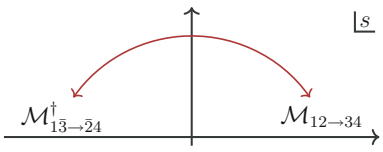
the Hermitian conjugated (anti-causal) scattering amplitude

$$\langle 0|a_4 a_3 b_2^\dagger b_1^\dagger|0\rangle = \langle 0|a_4 a_3 S^\dagger a_2^\dagger a_1^\dagger|0\rangle = \langle 34|\, S^\dagger\, |12\rangle \,, \tag{5.1.15}$$

and the forwards terms

$$\langle 0|a_4 a_3 a_2^\dagger a_1^\dagger|0\rangle = \langle 34|\, \mathbb{1}\, |12\rangle = \delta_{12,34} = \langle 0|b_4 b_3 b_2^\dagger b_1^\dagger|0\rangle \,. \tag{5.1.16}$$

Restricting two non-forward kinematics, we find only *two* four-point observables: (5.1.14) and (5.1.15). Given the age of this subject, it is reassuring that we did not discover anything new at four points. The fact that there are two observables and not just a single one is significant. As we will elaborate on later, crossing symmetry relates these two objects—amplitudes to their Hermitian conjugate, and vice versa!



(5.1.17)

Before moving on, let us stress that crossing is not simply a matter of "flipping the sign of energies in formulas". Consider for example the integrated expression for an s-channel massless bubble integral in minimal subtraction:

$$\text{Bub}(s) \propto \begin{cases} \log \frac{-s}{\mu^2} & s < 0 \\ \log \frac{s}{\mu^2} - i\pi & s > 0 \end{cases} \,. \tag{5.1.18}$$

To go from one expression to the other, we have to follow a precise *continuation path*. The arc in the upper-half s-plane tells us exactly the correct branch choice.

5.1.2.3 Enumerating Observables: *Five-Point*

As should now be clear, the task of cataloging all possible observables is a combinatorial exercise. Hence, as the number of points increases, we should encounter increasingly non-trivial possibilities, leading to more exotic observables beyond the conventional scattering amplitudes. This becomes manifest already for five points. For example, while we still have the $3 \leftarrow 2$ scattering amplitudes

$$\langle 0|b_5 b_4 b_3 a_2^\dagger a_1^\dagger|0\rangle = \langle 345|\, S\, |12\rangle \ , \tag{5.1.19}$$

we also have observables of the form

$$\langle 0|a_5 a_4 b_3 a_2^\dagger a_1^\dagger|0\rangle = \langle 45|\, b_3\, |12\rangle = \langle 45|\, S^\dagger a_3 S\, |12\rangle \ . \tag{5.1.20}$$

As the writing makes manifest, (5.1.20) computes an expectation value for observing particle 3 in a particular *in* state that contains two particles, and is totally agnostic as to what happens to these particles *after* particle 3 is emitted. (Both the bra and ket are *in* states: no future boundary conditions are imposed.) We can make (5.1.20) look more like a conventional quantum-mechanical expectation value by integrating the on-shell momenta p_1, p_2 against a wavepacket:

$$\begin{aligned} |\psi\rangle_{\text{in}} &\equiv \int_{p_1,p_2} \psi(p_1,p_2)\,|12\rangle \\ \Rightarrow\ {}_{\text{in}}\langle\psi|\, b_3\, |\psi\rangle_{\text{in}} &= \int_{p_1,p_2,p_4,p_5} \psi^*(p_4,p_5)\psi(p_1,p_2)\,\langle 45|\, b_3\, |12\rangle \ . \end{aligned} \tag{5.1.21}$$

Note that (5.1.21) is *not* measuring the type 3 particle number; such an observable would be quadratic in field 3. Instead, (5.1.21) measures the linear (leading order) response of the scattering background in field 3. In which realistic situations could this arise? For example, (5.1.21) could represent (1) the expectation value of the electromagnetic field associated with radiation after the collision of two charged particles (further discussed in Sect. 5.2.3), or (2) the expectation value of the graviton field (gravitational wave signal) generated by the scattering of two massive objects (e.g., black holes), which is further discussed in Sect. 5.1.3. Both are examples of radiative *waveforms*. In this context, the wavefunction ψ sets the initial energy and impact parameter of the colliding bodies (see [6]).

Let us reiterate a basic but important point: the signal observed at a detector such as LIGO/VIRGO/KAGRA is *linear* in metric perturbations. Linear measurements of electric fields, while not necessarily important nor natural at colliders, are familiar in low-energy physics: this is what a voltmeter measures. Although the definition (5.1.21) makes no reference to coherence effects, such linear measurements are typically only practical when the final state contains a coherent state of many photons or gravitons that makes the expectation value (5.1.21) large.

That said, it may not yet be clear from (5.1.21) *how* the observable should be computed in practice (e.g., using perturbation theory). To get started, it is useful to

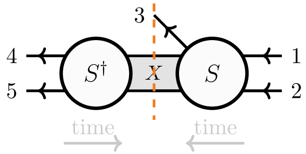
insert in (5.1.20) a complete basis of on-shell states

$$\mathbb{1} = \sum\!\!\!\!\!\!\!\int_X |X\rangle_{\text{in in}}\langle X| = \sum\!\!\!\!\!\!\!\int_X |X\rangle_{\text{out out}}\langle X| \ , \tag{5.1.22}$$

as follows:

$$\langle 0|a_5 a_4 b_3 a_2^\dagger a_1^\dagger|0\rangle = \sum\!\!\!\!\!\!\!\int_X \langle 45|\, S^\dagger \,|X\rangle \,\langle X3|\, S\,|12\rangle$$

$$= \quad .$$



(5.1.23)

To be clear, (5.1.22) is a sum over the complete basis of states $|X\rangle\,\langle X|$ integrated over the *on-shell phase space*

$$\sum\!\!\!\!\!\!\!\int_X |X\rangle\,\langle X| = \int \frac{\mathrm{d}^{\mathrm{D}-1} p_1}{(2\pi)^{\mathrm{D}-1} 2E_{p_1}}\, |p_1\rangle\,\langle p_1|$$

$$+ \int \prod_{j=1}^{2} \frac{\mathrm{d}^{\mathrm{D}-1} p_j}{(2\pi)^{\mathrm{D}-1} 2E_{p_j}} \frac{|p_1 p_2\rangle\langle p_1 p_2|}{2!} + \ldots \tag{5.1.24}$$

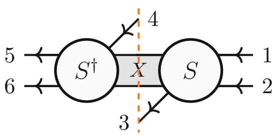
Now, from (5.1.23), one way to practically compute the waveform observable becomes apparent: it is the product of S-matrix elements (each computable separately using e.g. Feynman diagrams) connected by an on-shell phase-space integral. Note that this calculation is completely inclusive in the final state X: as far as we care, the colliding bodies could annihilate each other and we would still get a non-trivial waveform!

All other observables at five-point are either forward or obtained by Hermitian conjugation or time reversal of those listed above. Thus, we move on to six-point.

5.1.2.4 Enumerating Observables: *Six-Point*

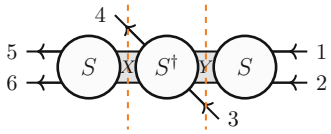
Similar to the previous instances, conventional six-point scattering amplitudes (i.e., $\langle 3456|\, S\,|12\rangle$, $\langle 456|\, S\,|123\rangle$, relevant for scattering of one parton against two partons and their Hermitian conjugates) remain. More intriguingly, we have observables of the form

$$\langle 0|a_6 a_5 b_4^\dagger b_3 a_2^\dagger a_1^\dagger|0\rangle = \quad ,$$



(5.1.25)

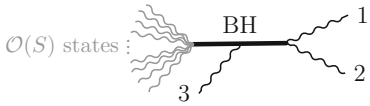
$$\langle 0|a_6 b_5 a_4 b_3^\dagger a_2^\dagger b_1^\dagger|0\rangle = \quad .$$



(5.1.26)

In the forward limit $(6, 5, 4 \to 1, 2, 3)$, the observable (5.1.25) is well known: it is the inclusive cross-section measuring the number of particles of type 3 produced in the scattering of particles 1 and 2: $\langle 0|a_6 a_5 b_4^\dagger b_3 a_2^\dagger a_1^\dagger|0\rangle \to \langle 12|\, N_3\, |12\rangle$, with $N_3 = b_3^\dagger b_3$ quadratic in the field.

The same observable (5.1.25) also appears in more exotic situations. For center-of-mass energies well above the Planck mass $\sqrt{s_{12}} \gg M_{\rm pl}$, one could imagine (5.1.25) as measuring the "square" of the following process:



(5.1.26)

where particles 1 and 2 collide and form a black hole, which eventually fully evaporates by emitting Hawking radiation. In a strict S-matrix approach, the exact final state is a complicated combination of $\mathcal{O}(S)$ quanta where S is the entropy of the intermediate black hole, with matrix elements that are typically exponentially small $\sim e^{-S/2}$ by general statistical considerations. The calculation of this quantum state remains an unsolved task, even in principle (although there was notable recent progress in understanding its statistical properties using replica wormholes). However, (5.1.25) is a much simpler object because it is agnostic about the complicated cloud of on-shell final states and only cares about the particular Hawking quantum labeled 3 in the figure. (By making the momenta p_3 and p_4 slightly non-forward and performing a Wigner transform, the observable is also sensitive to the arrival time of the measured radiation.) Viewing (5.1.25) as a two-point function in a black hole geometry also makes it clear that it can be reliably calculated without knowing the exact final state—in fact this is exactly how Hawking [7] calculated it!

Let us finally comment on (5.1.26), which is at face value a "maximal scrambling" of a's and b's. It is useful to consider particles 1 and 6 as defining a background state and to interpret this observable as a *four-point* OTOC (commutator squared) in this background. By transforming to the time domain and applying a large time translation to $b_3^\dagger$ and b_5, this process potentially probes the Lyapunov exponent characterizing the chaotic growth of small perturbation $b_3^\dagger$ (see [1] for more references). For a more complete list of six-point observables, see [2].

5.1.3 (Some) Crossing Moves

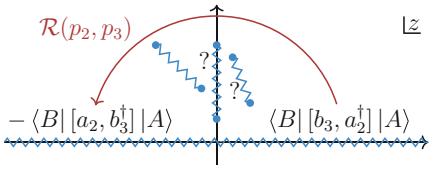
We now sketch how crossing symmetry bridges some of the above observables through analytic continuations in the kinematic space. The basic object (further discussed in Sect. 5.2.1) is the retarded product of currents $j \propto (-\partial^2 + m^2)\phi$,

$$\mathcal{R}(p_2, p_3) = \int \mathrm{d}^{\mathrm{D}}x_2 \mathrm{d}^{\mathrm{D}}x_3 \mathrm{e}^{-i(p_2\cdot x_2 + p_3\cdot x_3)} {}_{\mathrm{out}}\langle B|\, R_3(j_3 j_2)\, |A\rangle \ , \tag{5.1.27}$$

where $R_3(j_3 j_2) = [j_3, j_2]\theta(x_3^0 > x_2^0)$ denotes the retarded product. The current j amputates the external propagators and following the logic of the LSZ reduction formula one can show that for real on-shell momenta:

$$\mathcal{R}(p_2, p_3) = \begin{bmatrix} \langle B|\, [b_3, a_2^\dagger]\, |A\rangle & p_2^0 < 0\,,\ p_3^0 > 0\,, \\ -\langle B|\, [a_2, b_3^\dagger]\, |A\rangle & p_2^0 > 0\,,\ p_3^0 < 0\,. \end{bmatrix} \quad \text{(on-shell)} \tag{5.1.28}$$

The challenge of crossing is to parameterize a *continuous* path of analytic continuation that (1) is wholly supported on the mass shell and (2) connects the two preceding real configurations (5.1.28):



(5.1.29)

In the case of gapped (and infrared-trivial) quantum field theories, such a path is guaranteed to exist for $2 \leftarrow 2$ and $3 \leftarrow 2$ scattering thanks to the work of Bros, Epstein and Glaser initiated in the 1960s [8–11]. In gapless theories and at higher multiplicities, the existence of such a path remains conjectural to this day, although recent progress in this direction has been made [3,12]. The z-path we will consider is simply a complex boost applied to two of the external momenta; see [3] for details.

To develop some intuition about why the mass gap plays an important role without going into the technical details of [8–11], we can consider the following example.

Example 5.1.1 (Intuition for Analyticity) Under the change of variables $x_3 \mapsto x_2 + \Delta x$, (5.1.27) becomes

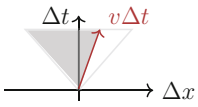
$$\mathcal{R}(p_2, p_3) = \int_{\Delta x \in \bar{V}^+} \mathrm{d}^{\mathrm{D}}\Delta x\, \mathrm{e}^{-ip_3\cdot\Delta x} \int \mathrm{d}^{\mathrm{D}}x_3 \mathrm{e}^{-i(p_2+p_3)\cdot x_2} {}_{\mathrm{out}}\langle B|\, R_3(j_3 j_2)\, |A\rangle \ . \tag{5.1.30}$$

The Fourier transform converges provided that (1) $\mathrm{Re}(-ip_3 \cdot \Delta x) \leqslant 0\ \forall \Delta x$ and (2) $(p_2 + p_3)^\mu \in \mathbb{R}^{1,3}$. Since the retarded product is supported in the (closure of) the forward lightcone $\Delta x \in \bar{V}^+$, we conclude that the integral is analytic for $\mathrm{Im}\, p_3^\mu \in V^+$ since then the Fourier transform converges exponentially. This is the basic reason why we expect analyticity in some upper-half plane.

The difficulty is that a timelike imaginary part for p_3^μ is incompatible with the mass-shell condition $p_3^2 + m^2 = 0$. Consider for example the on-shell momentum

$$p_3^\mu = E(1, v, 0, 0) + i\varepsilon(v, 1, 0, 0) \qquad (0 < v < 1 \,\text{real},\ E \gg \varepsilon > 0)\,. \tag{5.1.31}$$

The imaginary part is spacelike and point (1) is only satisfied if $\Delta x^1 < v\Delta t$, where $\Delta t = (x_3 - x_2)^0 > 0$ and $\Delta x^1 = (x_3 - x_2)^1$. Thus, we have a decaying exponential only for displacements Δx *inside* the shaded region:



(5.1.32)

We see that while we obtain analyticity in V^+ just from the causal support $|\Delta x^1| < \Delta t$ of the retarded product, analyticity on-shell will only hold if the retarded product is supported within the smaller shaded region $|\Delta x^1| < v\Delta t$. Physically, this is a constraint on the velocity of the intermediate particles produced.

This naive analysis suggests that analyticity is safe for sufficiently boosted external momenta in gapped theories, but can fail in gapless theories whenever *external* particles move more slowly than internal ones.

We hope that this intuition can be formalized in the near future using the concept of *essential support*, described in Sect. 5.3 below. For now, it remains unproven whether the domain of analyticity of (5.1.27) intersects with the mass shell, in general. We will proceed under the assumption that it does.

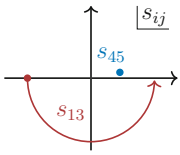
Example 5.1.2 (Five-Point Tree-Level with Massive Resonances) We now borrow a five-point example from [3] to illustrate how the inclusive observable (5.1.23) (equivalent to a product of conventional (time-ordered) amplitudes) is related to a five-point scattering amplitude through analytical continuation. Here, we work at tree level, which is (perhaps surprisingly) sufficient to highlight the non-trivial conceptual aspects of this exercise without the unnecessary complications arising from branch cuts.

Consider a cubic theory where the following tree-level process involving a heavy intermediate particle of mass M is possible

$$i\mathcal{M}^{\text{start}}_{345\leftarrow 12} = \quad = \frac{-ig^3}{(-s_{45}+M^2-i\varepsilon)(-s_{13}+M^2)}\,. \tag{5.1.33}$$

Note that the $i\varepsilon$ only matters if it is kinematically possible to produce M, in which case it is necessarily unstable ("if it can be produced, it can decay"). The tree-level diagrams here should be understood as a narrow-width approximation to this resonance.

In (5.1.33) we dropped the $i\varepsilon$ prescription for s_{13} since this invariant is spacelike ($s_{13} < 0$) in the kinematics under consideration and therefore its propagator cannot go on shell. Nevertheless, the $i\varepsilon$ must be retained for the timelike Mandelstam invariant s_{45}. We proceed now to analytically continue to the kinematic region where the energies of particles 2 and 3 flip signs. The invariant s_{13}, initially negative, rotates in the counterclockwise direction, ending its journey below the positive real axis, while the invariant s_{45} remains unchanged:



(5.1.34)

The result of this analytic continuation is written as

$$[i\mathcal{M}_{345\leftarrow 12}]_{\circlearrowleft s_{13}} = \frac{-ig^3}{(-s_{45}+M^2-i\varepsilon)(-s_{13}+M^2+i\varepsilon)}\,, \tag{5.1.35}$$

where now $s_{13} > 0$ and its $-i\varepsilon$ prescription matters. Observe that the second propagator acquires the "wrong" $i\varepsilon$ prescription. This detail is important because after the crossing we are now in $245 \leftarrow 13$ kinematics where this pole is accessible. The fact that s_{13} is on the "wrong side" of its cut mirrors what happens for $2 \leftarrow 2$ scattering, where the large arc (see (5.1.29)) lands us on the wrong side of the u-channel cut. However, here the s_{45} pole remained on the correct side, preventing identifying the right-hand side with a complex conjugated amplitude. *It is a different object*!

Our claim now is that the blob pattern in (5.1.23) is a shortcut to produce the same result as the analytic continuation we just went through. To see this, substitute $S^\dagger = \mathbb{1} - iT^\dagger$ and $S = \mathbb{1} + iT \to iT$ into (5.1.23); for S we could drop the $\mathbb{1}$ part since 1 and 2 must interact in order that 3 be emitted, but for $S^\dagger$ we have both

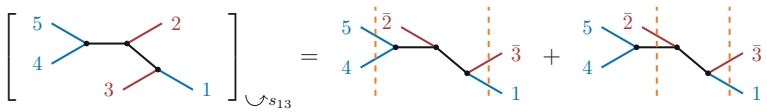
a disconnected and connected term. They give respectively the complex-conjugated five-point amplitude, plus a non-linear term where the heavy particle M is produced:

$$[i\mathcal{M}_{345\leftarrow 12}]_{\curvearrowright s_{13}} \stackrel{?}{=} i\mathcal{M}^{\dagger}_{245\leftarrow 13} + [i\mathcal{M}_{45\leftarrow M}]\, 2\pi\delta(-s_{45}+M^2)\, [i\mathcal{M}^{\dagger}_{M2\leftarrow 13}] \tag{5.1.36a}$$

$$= \frac{-ig^3}{(-s_{45}+M^2+i\varepsilon)(-s_{13}+M^2+i\varepsilon)} + \frac{2\pi\delta(-s_{45}+M^2)g^3}{-s_{13}+M^2+i\varepsilon} \tag{5.1.36b}$$

$$= \left(\frac{1}{-s_{45}+M^2+i\varepsilon} + 2\pi i\delta(-s_{45}+M^2)\right)\frac{-ig^3}{-s_{13}+M^2+i\varepsilon}\,. \tag{5.1.36c}$$

Using the identity $\frac{1}{x\pm i\varepsilon} = \mathrm{PV}\frac{1}{x} \mp i\pi\delta(x)$, the result manifestly agrees with (5.1.35)! Diagrammatically, we can summarize the discussion as follows:



(5.1.37)

Similar computations at loop level are discussed in [3], where a general diagrammatic statement of (5.1.29) is also given. In particular, we found that the type of computation described above worked for purely massless scattering (without a mass gap). However, we also point out counterexamples in cases where external particles move slower than internal particles, verifying the intuition developed below (5.1.32).

In the next lecture, we will show how the observables introduced above follow from the reduction of (O)TO products of on-shell currents.

5.2 Analyticity and IR Properties of Timefolded Observables

Mathieu Giroux

The purpose of this lecture is to revisit the application of reduction formulae for observables with unconventional time ordering through explicit examples. As motivation, we quickly revisit the so-called Lehmann–Symanzik–Zimmermann (LSZ) reduction formula [13]—which relates conventional time-ordered scattering amplitudes to time-ordered correlators—from a modern perspective. Later on,

we generalize this concept to out-of-time order correlators, which reduce to the aforementioned out-of-time order (OTO) observables.

5.2.1 LSZ Reduction Revisited

The idea behind the LSZ reduction formula is simple: it makes precise the idea that *time-ordered products of fields* encode more information than is needed to describe in-out particle collisions characterized by *scattering amplitudes*, namely

$$ {}_{\text{in}}\langle (j{+}1)\dots n|\, S\,|1\dots j\rangle_{\text{in}} = \Big[\prod_{k=1}^{n} i\int \mathrm{d}^{\mathrm{D}}x_k\, \mathrm{e}^{-ip_k\cdot x_k}\big[-\partial_{x_k}^2{+}m_k^2\big]\Big] \langle 0|\, T(\phi_1\phi_2\dots\phi_n)\,|0\rangle\,. \tag{5.2.1}$$

Indeed, because $\big[-\partial_x^2{+}m^2\big]\phi(x) = 0$ for asymptotic (free) fields, the physical content of (5.2.1) is that the S-matrix *projects out* (or "amputates") such fields from the time-ordered product. This means that only terms with a pole of the form $p^2+m^2=0$ in momentum space ultimately contribute to the scattering amplitude. For what follows, it is useful to revisit *why* (5.2.1) holds from a modern perspective.

The first thing we do is to express the right-hand side of (5.2.1) in terms of the currents defined by

$$T(j(x)\cdots) \equiv i(-\partial_x^2+m^2)T(\phi(x)\cdots)\,. \tag{5.2.2}$$

Given this definition, the right-hand side of (5.2.1) reduces to the Fourier transformed (momentum space) time-ordered correlator of currents

$$\begin{aligned}\text{RHS of (5.2.1)} &= \Big[\prod_{k=1}^{n}\int \mathrm{d}^{\mathrm{D}}x_k\, \mathrm{e}^{-ip_k\cdot x_k}\Big]\langle 0|\, T(j_1 j_2\dots j_n)\,|0\rangle \\ &= \langle 0|\, T(j(p_1)j(p_2)\dots j(p_n))\,|0\rangle\,.\end{aligned} \tag{5.2.3}$$

From there, the key observation is that the on-shell limit of the currents is a total derivative. In fact, an elementary computation shows that

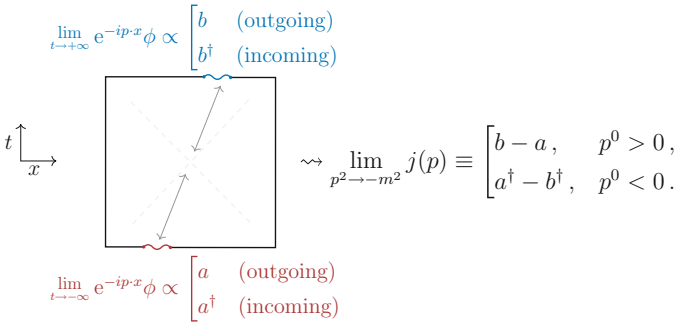
$$j(p)\equiv\int \mathrm{d}^{\mathrm{D}}x\,\mathrm{e}^{-ip\cdot x}\,j(x) \xrightarrow{\substack{\text{on-shell:}\\ p^2\to -m^2}} \int \mathrm{d}^{\mathrm{D}}x\,\frac{\partial}{\partial x^\mu}\Big[\mathrm{e}^{-ip\cdot x}\,(-i\partial_x^\mu+p^\mu)\phi(x)\Big]\,. \tag{5.2.4}$$

Now, we observe that the Fourier phase $\mathrm{e}^{-ip\cdot x}$ oscillates rapidly as $x\to\infty$, except possibly along the direction of the external particles: $x^\mu\propto\pm p^\mu$. Along such directions, particles are eventually detected by a detector infinitely far in the

future (or past), which is only possible if *constructive interference* occurs between the rapidly oscillatory Fourier phase and the phase of the field

$$\lim_{t\to\pm\infty} \mathrm{e}^{-ip\cdot x}\phi = non-\text{oscillatory quantity along } x^\mu \propto \pm p^\mu \,. \tag{5.2.5}$$

In other words, (5.2.5) is the minimal condition that ensures non-zero asymptotic measurements ($j(p) \neq 0$). Thus, the on-shell current in (5.2.4) reduces to non-trivial surface terms (which *defines* for us the a, b, $a^\dagger$ and $b^\dagger$ introduced earlier) *only* along the trajectory of the particle. This is schematically summarized as follows:



(5.2.6)

The cartoon on the left illustrates the asymptotic measurements made in the far past and future ($a, a^\dagger$ and $b, b^\dagger$, respectively) and emphasizes that they follow from constructive interference (see (5.2.5)) at late times along the direction of particles; along other directions, destructive interference occurs and no measurement is made. (The dotted lines show light-cone axes centered around the collision region at finite time.) The sign of p^0 determines which case in (5.2.6) is applicable.

The LSZ reduction formula of a time-ordered correlator easily follows: apply (5.2.6) to each current then follow the T-ordering instruction to bring all b and $b^\dagger$ to the left of the a and $a^\dagger$ (assuming here that $p_i^0 < 0$ for $i \leqslant j$)

$$\begin{aligned}\langle 0|\, T(j_1(p_1)j_2(p_2)\dots j_n(p_n))\,|0\rangle &= \langle 0|\, T(a_1^\dagger\cdots a_j^\dagger b_{j+1}\cdots b_n)\,|0\rangle \\ &= \langle 0|\, b_{j+1}\cdots b_n a_1^\dagger\cdots a_j^\dagger\,|0\rangle \\ &= {}_{\text{in}}\langle (j{+}1)\dots n|\, S\,|1\dots j\rangle_{\text{in}}\,.\end{aligned} \tag{5.2.7}$$

Here we have also used the fact that $a|0\rangle = 0 = \langle 0|\, b^\dagger$ to eliminate all terms with an a or $b^\dagger$ in (5.2.6). Note that this argument (and the LSZ reduction formula itself, as far as we understand) requires non-forward kinematics: $(p_i + p_j)^\mu \neq 0$ for each pair.

For out-of-time-ordered observables (which we study next), we must generally keep all the terms in (5.2.6). This leads to generalized LSZ reduction formulae!

5.2.2 Generalized Reduction Formulae

As reviewed above, reduction formulae connect scattering amplitudes with the on-shell limit of amputated Green functions. Below, we illustrate how generalized reduction formulae can be used to reveal non-trivial abstract features of either time-ordered or out-of-time-ordered (OTO) observables (such as analyticity). In this chapter, we examine how applying LSZ-like reductions to generic OTO correlators (OTOC) reveals a wide range of asymptotic observables. We do so from two perspectives: from an algebraic point of view (involving formal manipulations of various time ordering operators) and visually using the Schwinger–Keldysh formalism.

5.2.2.1 OTOC: Definitions and Generalities

The type of correlators we wish to reduce will always be expressible in terms of products of *time-ordered products* of the following form (introduced in the 1960s by Ruelle and Araki–Burgoyne [14, 15])

$$\langle 0|T(\phi_1 \dots \phi_{i_1})T(\phi_{i_1+1} \dots \phi_{i_2}) \dots T(\phi_{i_{n-1}+1} \dots \phi_{i_n})|0\rangle \,, \tag{5.2.8}$$

where $\phi_n = \phi(x_n)$ is a local operator (field)[1] and $T(\phi_1 \dots \phi_n)$ is the conventional time-ordered product of n local operators

$$\begin{aligned} T(\phi_1 \dots \phi_n) &= \sum_{\sigma \in \mathrm{S}_n} \theta(x^0_{\sigma(1)} > x^0_{\sigma(2)} > \dots > x^0_{\sigma(n)}) \epsilon(\sigma) \phi_{\sigma(1)} \phi_{\sigma(2)} \dots \phi_{\sigma(n)} \\ &= \sum_{\sigma \in \mathrm{S}_n} \Big[\prod_{m=1}^{n-1} \theta(x^0_{\sigma(m)} > x^0_{\sigma(m+1)}) \Big] \epsilon(\sigma) \phi_{\sigma(1)} \phi_{\sigma(2)} \dots \phi_{\sigma(n)} \,, \end{aligned} \tag{5.2.9}$$

where S_n denotes the symmetric group of degree n and

$$\epsilon(\sigma) = \begin{bmatrix} 1 & \text{bosonic operators} \\ \mathrm{sign}(\sigma) & \text{fermionic operators} \end{bmatrix} . \tag{5.2.10}$$

For example, for the bosonic two-point function, we get the familiar expression

$$T(\phi_1\phi_2) = \phi_1\phi_2\theta(x^0_1 > x^0_2) + \phi_2\phi_1\theta(x^0_2 > x^0_1) \,. \tag{5.2.11}$$

Interestingly, linear combinations of products of T's span other useful and physical objects such as *anti-time-ordered products* $\overline{T}(\dots)$, *retarded products* $R_i(\dots)$ and *advanced products* $A_i(\dots)$. At n-point, the former is simply defined by reversing

[1] Note that the notation is abused here; the fields are allowed to be of different nature, which is not apparent as we suppressed this label.

the inequalities in (5.2.9), while the second and third are defined as follows

$$R_0(\phi_0\phi_1\dots\phi_n) = \sum_{\sigma\in S_n} \theta(x_0^0{>}x_{\sigma(1)}^0{>}\dots{>}x_{\sigma(n)}^0) \\ \underbrace{\big[\big[\dots\big[\big[}_{(n-1)\text{-fold}}\phi_0,\phi_{\sigma(1)}\big],\phi_{\sigma(2)}\big],\dots\big],\phi_{\sigma(n)}\big]\,, \tag{5.2.12a}$$

$$A_0(\phi_0\phi_1\dots\phi_n) = \sum_{\sigma\in S_n} \theta(x_0^0{<}x_{\sigma(1)}^0{<}\dots{<}x_{\sigma(n)}^0) \\ \big[\phi_{\sigma(n)},\big[\dots,\big[\phi_{\sigma(2)},\big[\phi_{\sigma(1)},\underbrace{\phi_0\big]\big]\dots\big]\big]}_{(n-1)\text{-fold}}\,, \tag{5.2.12b}$$

with the special case $R_0(\phi_0) = \phi_0$. The A-product is like the R product with commutators nested in the second entry rather than in the first one, and with the inequalities in the θ-functions reversed.

Importantly, the R-product (as well as the A-product) is symmetric in $\{1, 2, \dots, n\}$, but one field ϕ_0 is singled out as the one with the latest time; it is *pinned* in the future of everyone. That is,

$$R_0(\phi_0\phi_1\dots\phi_n) = 0 \qquad \text{if } x_0^0 < \max(x_1^0, x_2^0, \dots, x_n^0)\,. \tag{5.2.13}$$

Example 5.2.1 To digest these definitions, let us look at some low-n examples:

$$R_2(\phi_2\phi_1) = R_2(\phi_1\phi_2) = [\phi_2,\phi_1]\theta(x_2^0 > x_1^0)\,, \\ R_3(\phi_3\phi_2\phi_1) = R_3(\phi_1\phi_2\phi_3) = [[\phi_3,\phi_2],\phi_1]\theta(x_3^0 > x_2^0)\theta(x_2^0 > x_1^0) + (1 \leftrightarrow 2)\,. \tag{5.2.14}$$

Note that both (5.2.9) and (5.2.12a) involve the sum over $(n-1)!$ terms. Above, we indicated that both the time reversal of (5.2.9) and (5.2.12a) can be expressed solely in terms of standard time-ordered products. This extends to any number of fields, as we can show by a θ-functions bookkeeping. Examining the simple case of $n = 2$ is sufficient to understand why. Indeed, introducing the shorthand notation

$$\theta_{ij} = \theta(x_i^0 > x_j^0)\,, \tag{5.2.15}$$

we have

$$\begin{aligned} \overline{T}(\phi_1\phi_2) &= \phi_1\phi_2\theta_{21} + \phi_2\phi_1\theta_{12} \\ &= \phi_1\phi_2(\theta_{21}+\theta_{12}) - \phi_1\phi_2\theta_{12} + \phi_2\phi_1(\theta_{12}+\theta_{21}) - \phi_2\phi_1\theta_{21} \\ &= \phi_1\phi_2 - \phi_1\phi_2\theta_{12} + \phi_2\phi_1 - \phi_2\phi_1\theta_{21} \\ &= T(\phi_1)T(\phi_2) + T(\phi_2)T(\phi_1) - T(\phi_1\phi_2)\,, \end{aligned} \tag{5.2.16}$$

and

$$\begin{aligned} R_1(\phi_1\phi_2) &= [\phi_1, \phi_2]\theta_{12} = \phi_1\phi_2\theta_{12} - \phi_2\phi_1\theta_{12} \\ &= \phi_1\phi_2\theta_{12} - \phi_2\phi_1(\theta_{12} + \theta_{21}) + \phi_2\phi_1\theta_{21} \\ &= T(\phi_1\phi_2) - T(\phi_2)T(\phi_1)\,. \end{aligned} \tag{5.2.17}$$

We emphasize once more that such relations generalize to any number of points and are invertible. Consequently, replacing T by $\overline{T}$ or R in (5.2.8) would generate the same linear span of correlators.

In the following, we revisit the Schwinger–Keldysh formalism, as described in the contribution [16]. This formalism provides a useful visual approach to understanding OTOCs.

5.2.2.2 Schwinger–Keldysh Formalism

In order to better understand OTOCs and their evaluation in many situations, it is useful to turn to a path-integral representation. This is precisely what *Schwinger–Keldysh formalism* (SK) provides. The idea behind this formalism is quite simple:

This approach allows us to write down any combination of anti- and time-ordered products by linking the corresponding Lorentzian (real) time axes by infinitesimal Euclidean (imaginary) time shifts. Ultimately, this results in a path integral that features numerous time-folds that oscillate between the past and the future.

Example 5.2.2 To illustrate this in the context of correlators, we can consider the following seven-field correlator represented by a path integral over a Schwinger–Keldysh contour with three timefolds (I, II, and III):

$$\langle 0|\underbrace{T(FG)}_{\text{fold III}}\underbrace{\overline{T}(CDE)}_{\text{fold II}}\underbrace{T(AB)}_{\text{fold I}}|0\rangle \equiv \langle 0|\mathcal{C}(A^{\mathrm{I}}B^{\mathrm{I}}C^{\mathrm{II}}D^{\mathrm{II}}E^{\mathrm{II}}F^{\mathrm{III}}G^{\mathrm{III}})|0\rangle$$

$$= \quad \begin{array}{l} \mathrm{I} \\ \mathrm{II} \\ \mathrm{III} \end{array} \quad \begin{array}{l} \text{B} \quad \text{A} \\ \text{C} \quad \text{D} \quad \text{E} \\ \text{G} \quad \text{F} \end{array} \quad \begin{array}{l} \mathrm{Im}\, t \\ \mathrm{Re}\, t \end{array} \,. \tag{5.2.18}$$

The *contour-ordering symbol* $\mathcal{C}$ exemplified here is a natural generalization of the time-ordering symbol. The superscript on an operator labels on which timefold it is inserted. The total action includes the contribution from past directed branches with a minus sign (in our convention, the even ones: II, IV, ...) since $\mathrm{d}t < 0$, such that $\mathrm{e}^{iS_{\mathcal{C}}} = \mathrm{e}^{iS^{\mathrm{I}} - iS^{\mathrm{II}} + iS^{\mathrm{III}} - iS^{\mathrm{IV}} + \cdots}$. The path integral measure is simply a product of the usual one: $\int \mathcal{D}\phi = \int \mathcal{D}\phi^{\mathrm{I}} \int \mathcal{D}\phi^{\mathrm{II}} \cdots$.

Example 5.2.3 (Path Integral for Retarded Products) Let us consider what kind of formula we might expect for retarded functions by considering the simplest case of $n = 2$:

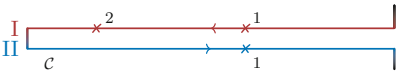
$$\langle 0|R_2(\phi_1\phi_2)|0\rangle = \langle 0|[\phi_2, \phi_1]\theta_{21}|0\rangle = \langle 0|(\phi_2\phi_1 - \phi_1\phi_2)\theta_{21}|0\rangle\,. \tag{5.2.19}$$

If we were to represent the two terms in parentheses, ignoring the θ-function, one would obtain the following pictures

(5.2.20)

However, it is clear that the second contour cannot accurately describe (5.2.19) precisely due to the presence of the θ function, which requires that 1 is in the past of 2. A straightforward and correct solution to this problem is to *superpose* both contours and to *fold* them as follows

$$\langle 0|R_2(\phi_1\phi_2)|0\rangle =$$



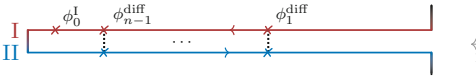
$$= \langle 0|\mathcal{C}(\phi_2(\phi_1^{\mathrm{I}} - \phi_1^{\mathrm{II}}))|0\rangle$$
$$\equiv \langle 0|\mathcal{C}(\phi_2\phi_1^{\mathrm{diff}})|0\rangle\,,$$

(5.2.21)

where $\phi_k^{\mathrm{diff}}(x) \equiv \phi_k^{\mathrm{I}}(x) - \phi_k^{\mathrm{II}}(x)$. Note that the minus sign comes from the commutator and that the role of the $\mathcal{C}$-operation is simply to bring any ϕ^{II} to the left of ϕ^{I}.

At n-point, we replace the commutator in (5.2.19) by the nested commutator in (5.2.12a) and repeat the same exercise to get the n-point retarded product in terms of the n-point correlators of field differences, namely

$$R_0(\phi_0\phi_1\cdots\phi_{n-1}) \equiv \mathcal{C}(\phi_0^{\mathrm{I}}\phi_1^{\mathrm{diff}}\cdots\phi_{n-1}^{\mathrm{diff}})$$
$$=$$



(5.2.22)

Note that the relative positions of the fields $1, \ldots, n-1$ with respect to each other on the timefolds are arbitrary; the key point is that ϕ_0 has to be the farthest in the future; otherwise we get zero. For advanced products, ϕ_0 is simply positioned in the past relative to all other fields.

Let us now illustrate how Cutkosky-like cutting rules arise from the Schwinger–Keldysh formalism.

Example 5.2.4 The *largest time equation* [17]

$$\sum_{k=0}^{n}(-1)^k \sum_{\sigma\in P(k,n-k)} \overline{T}(\phi_{\sigma(1)}\dots\phi_{\sigma(k)})T(\phi_{\sigma(k+1)}\dots\phi_{\sigma(n)}) = 0\,, \tag{5.2.23}$$

(where $P(k,n-k)$ denotes the set of partitions of n labels into two sets of size k and $n-k$) which reflects the fact that unitarity cuts are not independent of each other, takes a very simple form in the SK formalism

$$\langle 0|\,\mathcal{C}(\phi_1^{\text{diff}}\phi_2^{\text{diff}}\cdots\phi_n^{\text{diff}})\,|0\rangle = 0\,. \tag{5.2.24}$$

This equation is manifestly true thanks to the boundary condition

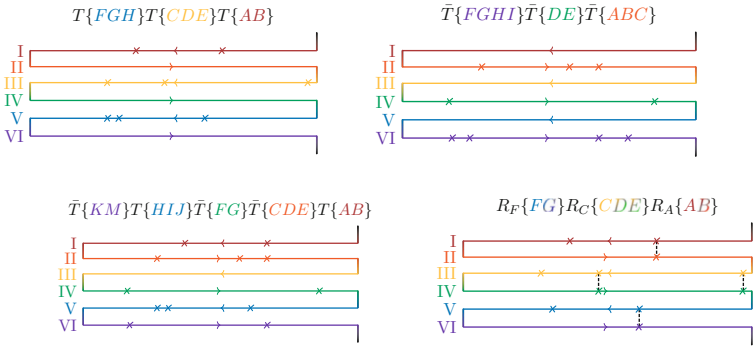
$$\lim_{x^0\to\infty} \phi^{\text{I}} - \phi^{\text{II}} = 0\,, \tag{5.2.25}$$

where the timefolds meet. It is not difficult to check that (5.2.24) is equivalent to (5.2.23). This is particularly easy to check explicitly on low n cases, e.g., for $n=2$

$$\begin{aligned}(5.2.23) \rightsquigarrow\ & (-1)^2 T(\phi_1\phi_2)+(-1)^2\overline{T}(\phi_1\phi_2)\\ &+(-1)^1\overline{T}(\phi_1)T(\phi_2)+(-1)^1\overline{T}(\phi_2)T(\phi_1) = 0\,,\\ (5.2.24) \rightsquigarrow\ & \langle 0|\,\phi_1^{\text{I}}\phi_2^{\text{I}}-\phi_2^{\text{II}}\phi_1^{\text{I}}-\phi_1^{\text{II}}\phi_2^{\text{I}}+\phi_1^{\text{II}}\phi_1^{\text{II}}\,|0\rangle = 0\,.\end{aligned} \tag{5.2.26}$$

We conclude our discussion of the Schwinger–Keldysh formalism's generalities by presenting some illustrative examples of how various products of T, $\bar{T}$ and R-products can be realized on multifold contours.

Example 5.2.5 Various types of time-ordering products discussed in the literature and their relations to Schwinger–Keldysh time-folds. For illustrative purposes, each panel only shows a three-time-fold contour, with the generalization to an arbitrary number of time-folds following the same obvious patterns:



(5.2.27)

On the top left, we have a product of time-ordered products T, while on the top right, we have a product of anti-time-ordered products $\bar{T}$. The cartoon on the bottom left represents a mixed product of time- and anti-time-ordered products, while the cartoon on the bottom right represents a product of retarded products R. As before, time goes from right to left in each panel.

The takeaway point here is that the set of operators spanned by multi-timefold Schwinger–Keldysh correlators such as (5.2.27) is, after θ-functions bookkeeping, in 1–1 correspondence with the set identified in (5.2.8). Next, we will extend the familiar LSZ reduction formula revisited earlier to the OTOC we just introduced.

5.2.2.3 Making Manifest Analyticity Properties

In what follows, we consider a few examples of interesting reduction formulae. As a warm-up, we consider once more conventional time-ordered amplitudes with the aim of establishing the so-called *optical theorem*/discontinuity/unitarity formula for $2 \leftarrow 2$ amplitudes. Later, applications at higher multiplicity will be discussed.

Example 5.2.6 (Four-Point Amplitudes and Discontinuities) In order to achieve our first goal, let us recall a useful implication of the stability condition

$$S a^\dagger |0\rangle = a^\dagger |0\rangle = b^\dagger |0\rangle = S b^\dagger |0\rangle \ , \tag{5.2.28}$$

following from the axioms introduced earlier, namely that

$$b_2^\dagger a_1^\dagger |0\rangle = b_1^\dagger a_2^\dagger |0\rangle = b_1^\dagger b_2^\dagger |0\rangle \ . \tag{5.2.29}$$

Using these, we can prove a useful intermediate result: in $34 \leftarrow 12$ kinematics, we have

$$\begin{aligned} T(j_1 j_2)|0\rangle &= T((a_1^\dagger - b_1^\dagger)(a_2^\dagger - b_2^\dagger))|0\rangle \\ &= T(a_1^\dagger a_2^\dagger - a_1^\dagger b_2^\dagger - b_1^\dagger a_2^\dagger + b_1^\dagger b_2^\dagger)|0\rangle \\ &= (a_1^\dagger a_2^\dagger - b_2^\dagger a_1^\dagger - b_1^\dagger a_2^\dagger + b_1^\dagger b_2^\dagger)|0\rangle \\ &= (a_1^\dagger a_2^\dagger - b_1^\dagger b_2^\dagger)|0\rangle \\ &= |12\rangle - S^\dagger a_1^\dagger a_2^\dagger S|0\rangle \\ &= (\mathbb{1} - S^\dagger)\,|12\rangle \ . \end{aligned} \tag{5.2.30}$$

We see from this simple calculation that $T(j_1 j_2)|0\rangle$ simply picks up the interacting part of S (it vanishes when the particles do not interact)! This identity can, of course, be applied to any state. For example, one can obtain the fully connected part of the

amplitude as follows

$$\begin{aligned} {}_{\text{out}}\langle 34|\, T(j_1 j_2)|0\rangle &= {}_{\text{in}}\langle 34|\, S(\mathbb{1} - S^\dagger)\,|12\rangle \\ &= {}_{\text{in}}\langle 34|\, (S - \mathbb{1})\,|12\rangle \\ &= i\mathcal{M}_{34\leftarrow 12} \end{aligned} \tag{5.2.31}$$

More interestingly, we can consider the correlator $\langle 0|\overline{T}(j_3 j_4)\mathrm{T}j_1 j_2|0\rangle$. On the one hand, we can insert a complete basis of asymptotic states via the resolution of identity

$$\mathbb{1} = \sum\!\!\!\!\!\!\int_X |X\rangle_{\text{in}}\ {}_{\text{in}}\langle X| = \sum\!\!\!\!\!\!\int_X |X\rangle_{\text{out}\,\text{out}}\,\langle X|\ , \tag{5.2.32}$$

according to

$$\langle 0|\overline{T}(j_3 j_4)\mathrm{T}j_1 j_2|0\rangle = \sum\!\!\!\!\!\!\int_X \langle 0|\overline{T}(j_3 j_4)\,|X\rangle_{\text{out}\,\text{out}}\,\langle X|\,\mathrm{T}j_1 j_2|0\rangle \tag{5.2.33a}$$

$$= \sum\!\!\!\!\!\!\int_X \langle 34|\,(\mathbb{1} - S)\,|X\rangle_{\text{out}\,\text{out}}\,\langle X|\,(\mathbb{1} - S^\dagger)\,|12\rangle \tag{5.2.33b}$$

$$= \sum\!\!\!\!\!\!\int_X \langle 34|\,(\mathbb{1} - S)S^\dagger\,|X\rangle_{\text{in}\,\text{in}}\langle X|\,S(\mathbb{1} - S^\dagger)\,|12\rangle \tag{5.2.33c}$$

$$= \sum\!\!\!\!\!\!\int_X \mathcal{M}^\dagger_{34\leftarrow X}\mathcal{M}_{X\leftarrow 12}\,. \tag{5.2.33d}$$

On the other hand, we can directly compute $\langle 0|\overline{T}(j_3 j_4)\mathrm{T}j_1 j_2|0\rangle$ using (5.2.30) twice:

$$\begin{aligned} \langle 0|\overline{T}(j_3 j_4)\mathrm{T}j_1 j_2|0\rangle &= (\langle 0|\overline{T}(j_3 j_4))(\mathrm{T}j_1 j_2|0\rangle) \\ &= (\langle 0|(a_3 a_4 - b_3 b_4))((a_1^\dagger a_2^\dagger - b_1^\dagger b_2^\dagger)|0\rangle)\,. \end{aligned} \tag{5.2.34}$$

Dropping the forward terms and using

$$(\mathcal{M}_{A\leftarrow B})^* = \langle A|S|B\rangle^* = \langle B|S^\dagger|A\rangle = \mathcal{M}^\dagger_{B\leftarrow A}\,, \tag{5.2.35}$$

we are left with

$$\begin{aligned} \langle 0|\overline{T}(j_3 j_4)\mathrm{T}j_1 j_2|0\rangle &= -\langle 0|a_3 a_4 b_1^\dagger b_2^\dagger|0\rangle - \langle 0|b_3 b_4 a_1^\dagger a_2^\dagger|0\rangle \\ &= -(\langle 0|b_1 b_2 a_3^\dagger a_4^\dagger|0\rangle)^\dagger - \langle 0|b_3 b_4 a_1^\dagger a_2^\dagger|0\rangle \\ &= i\mathcal{M}^\dagger_{34\leftarrow 12} - i\mathcal{M}_{34\leftarrow 12}\,. \end{aligned} \tag{5.2.36}$$

Comparing both sides of (5.2.33) and (5.2.36) gives the optical theorem.

Example 5.2.7 (The Role of R-products in Axiomatic Field Theory) Let us first illustrate why R-products are useful, again, with conventional time-ordered $2 \leftarrow 2$ scattering amplitudes. A first useful identity can be derived from the decomposition of the 2-point R-product in terms of T-products where both p_1 and p_2 have negative energies:

$$\begin{aligned} R_2(j_1 j_2) &= \mathrm{T} j_1 j_2 - \mathrm{T} j_1 \mathrm{T} j_2 \\ &= T((a_1^\dagger - b_1^\dagger)(a_2^\dagger - b_2^\dagger)) - (a_1^\dagger - b_1^\dagger)(a_2^\dagger - b_2^\dagger) \\ &= (\cancel{a_1^\dagger a_2^\dagger} - b_2^\dagger a_1^\dagger - \cancel{b_1^\dagger a_2^\dagger} + \cancel{b_1^\dagger b_2^\dagger}) - (\cancel{a_1^\dagger a_2^\dagger} - a_1^\dagger b_2^\dagger - \cancel{b_1^\dagger a_2^\dagger} + \cancel{b_1^\dagger b_2^\dagger}) \\ &= -[b_2^\dagger, a_1^\dagger]\,. \end{aligned} \tag{5.2.37}$$

Note that physically, the terms involving $b_1^\dagger$ were expected to cancel because j_1 must be in the past lightcone of j_2 and therefore it cannot be an "out" measurement. Consequently, particle 1 cannot reach future infinity (it returns to past infinity). From there, an elementary computation using (5.2.29) shows that

$$\begin{aligned} {}_{\text{out}}\langle 34|\, R_2(j_1 j_2)|0\rangle &= -{}_{\text{out}}\langle 34|\, [b_2^\dagger, a_1^\dagger]|0\rangle \\ &= -{}_{\text{out}}\langle 34|\, b_2^\dagger a_1^\dagger|0\rangle + {}_{\text{out}}\langle 34|\, a_1^\dagger b_2^\dagger|0\rangle \\ &= -\underbrace{{}_{\text{out}}\langle 34|\, b_2^\dagger b_1^\dagger|0\rangle}_{\text{Forward!}} + {}_{\text{out}}\langle 34|\, a_1^\dagger a_2^\dagger|0\rangle \\ &= {}_{\text{out}}\langle 34|\, a_1^\dagger a_2^\dagger|0\rangle = \langle 34|\, S\,|12\rangle\ . \end{aligned} \tag{5.2.38}$$

Thus, the conventional time-ordered scattering amplitude is also equal to the on-shell limit of some R products. What is interesting is that the R-product is supported over $x_2^0 > x_1^0$. This guarantees that its Fourier transform

$$\mathcal{R}(p_1, p_2) = \int \mathrm{d}^{\mathrm{D}} x_1 \mathrm{d}^{\mathrm{D}} x_2 \mathrm{e}^{-i(p_1\cdot x_1 + p_2 \cdot x_2)}\, {}_{\text{out}}\langle 34|\, R_2(j_1 j_2)|0\rangle\,, \tag{5.2.39}$$

is analytic/converges in a certain domain, in which equal positive timelike imaginary parts are added to p_2^μ and $-p_1^\mu$. Indeed, under the relabeling $x_1 = x_2 + y$, we have

$$\mathcal{R}(p_1, p_2) = \int \mathrm{d}^{\mathrm{D}} y \mathrm{e}^{-ip_1\cdot y} \int \mathrm{d}^{\mathrm{D}} x_2 \mathrm{e}^{-i(p_1+p_2)\cdot x_2}\, {}_{\text{out}}\langle 34|\, R_2(j_1 j_2)|0\rangle\,, \tag{5.2.40}$$

$$x_2^0 > x_1^0 \implies y^0 = x_1^0 - x_2^0 < 0 \implies y \in V^-\,. \tag{5.2.41}$$

Now, because $x_2^0 > x_1^0$, we must have $y^0 = x_1^0 - x_2^0 < 0$ and we conclude that y is in the past lightcone V^- (past pointing). This means that the Fourier transform only

converges provided that the first exponential decays, while the second oscillates. This happens precisely when

$$\mathrm{Re}(-iy\cdot p_1) = y\cdot \mathrm{Im}(p_1) < 0 \quad \text{while} \quad p_1+p_2 \in \mathbb{R}^{1,3}\,. \tag{5.2.42}$$

Since we work in mostly-plus signature, asking y and $\mathrm{Im}(p_1)$ to be timelike separated (implying that $\mathrm{Im}(p_1) \in V^-$) is equivalent to the first condition above. Thus, (5.2.38) indicates that the amplitude is *at least* analytic provided that equal and opposite timelike imaginary parts are added to p_1^μ and p_2^μ.

Of course, one could also represent the amplitude using the advanced product

$$A_2(j_1 j_2) = [j_1, j_2]\theta_{12} = -[j_2, j_1]\theta_{12} = -[j_2, j_1] + R_2(j_1 j_2)\,. \tag{5.2.43}$$

Indeed, since it differs from the retarded product only by a commutator (which actually vanishes in this kinematic) involving no θ-function, we have

$${}_{\text{out}}\langle 34|\, A_2(j_1 j_2)|0\rangle = {}_{\text{out}}\langle 34|\, R_2(j_1 j_2)|0\rangle - \underbrace{{}_{\text{out}}\langle 34|\,[j_2, j_1]|0\rangle}_{=0} = (5.2.38)\,. \tag{5.2.44}$$

Therefore, taking the Fourier transform of ${}_{\text{out}}\langle 34|\, A_2(j_1 j_2)|0\rangle$, one would obtain a similar set of conditions to that in (5.2.42), but with the inequality sign reversed. By "edge-of-the-wedge" type arguments, the equality of retarded and advanced representations suggests that the amplitude is analytic in a larger domain where equal and opposite imaginary parts are indeed added to p_1^μ and p_2^μ, but with the timelike requirement relaxed—as required to be on-shell.

The existence of multiple reduction formulae for the same quantity has significant implications. In particular, they are often central in axiomatic proofs that scattering amplitudes are analytic, for example, in a neighborhood of the mass shell [8, 9, 18]. The argument also generalizes to $m \leftarrow n$ scattering and has been used to demonstrate that amplitudes near the mass shell are essentially a finite sum of analytic functions [10], with a single function sufficing for $2 \leftarrow 2$ scattering. Moreover, while there exist only two distinct asymptotic observables at four points, namely $\mathcal{M}$ and $\mathcal{M}^\dagger$, these can be written in terms of a significantly greater number of correlation functions which enjoy distinct analyticity properties. We expect that a solid understanding of these correlation functions will add fruitful insight into the analytic properties of amplitudes.

Another reduction formula can be obtained similarly to (5.2.38) by inserting a retarded product between one in and out *single*-particle states. For example, focusing again on $34 \leftarrow 12$ kinematics, inserting

$$\begin{aligned} R_3\{j_3 j_2\} &= \mathrm{T} j_3 j_2 - \mathrm{T} j_2 \mathrm{T} j_3 \\ &= \mathrm{T}(b_3 - a_3)(a_2^\dagger - b_2^\dagger) - (a_2^\dagger - b_2^\dagger)(b_3 - a_3) \\ &= [b_3, a_2^\dagger]\,, \end{aligned} \tag{5.2.45}$$

we get (after using stability once more)

$$\langle 4|R_3\{j_3 j_2\}|1\rangle = \langle 4|b_3 a_2^\dagger|1\rangle = \langle 4|a_3 S a_2^\dagger|1\rangle = \langle 34|S|12\rangle\,. \tag{5.2.46}$$

This formula is relevant to analyticity at fixed-t and crossing symmetry through the upper-half s-plane [9]. When the energies of particles 2 and 3 are flipped

$$\begin{aligned} R_3\{j_3 j_2\} &= \mathrm{T} j_3 j_2 - \mathrm{T} j_2 \mathrm{T} j_3 \\ &= \mathrm{T}(a_3^\dagger - b_3^\dagger)(b_2 - a_2) - (b_2 - a_2)(a_3^\dagger - b_3^\dagger) \\ &= [b_3^\dagger, a_2]\,, \end{aligned} \tag{5.2.47}$$

for the same correlator, we get

$$\langle 4|R_3\{j_3 j_2\}|1\rangle = -\langle 4|a_2 b_3^\dagger|1\rangle = -\langle 4|a_2 S^\dagger a_3^\dagger|1\rangle = -\langle 42|S^\dagger|13\rangle\,. \tag{5.2.48}$$

Thus, it reduces to $\mathcal{M}^\dagger$ instead of $\mathcal{M}$. This is essentially the $2 \to 2$ case of *crossing symmetry* in (5.1.28) (see (5.1.17))! The general case of (5.1.28) is obtained in a similar way by not using the stability condition.

We now describe for the first time a higher point application of retarded products.

Example 5.2.8 (Higher point discontinuities and Steinmann relations) In $234 \leftarrow 12$ kinematics, consider the on-shell limit:

$$-\langle 0|R_5(j_5 j_4) R_3(j_3 j_2) a_1^\dagger|0\rangle = \ ? \tag{5.2.49}$$

Its reduction in terms of familiar time-ordered objects combines (5.2.37) (but without daggers) and (5.2.45). Indeed, using stability $\langle 0|\, a = \langle 0|\, b$ once again, we have

$$\begin{aligned} -\langle 0|R_5(j_5 j_4) R_3(j_3 j_2) a_1^\dagger|0\rangle &= -\langle 0|(-[b_5, a_4])([b_3, a_2^\dagger])a_1^\dagger|0\rangle \\ &= \langle 0|a_5 a_4 b_3 a_2^\dagger a_1^\dagger|0\rangle - \langle 0|b_4 b_5 b_3 a_2^\dagger a_1^\dagger|0\rangle \\ &= \langle 0|a_5 a_4 S^\dagger a_3 S a_2^\dagger a_1^\dagger|0\rangle - \langle 0|a_4 a_5 a_3 S a_2^\dagger a_1^\dagger|0\rangle\,. \end{aligned} \tag{5.2.50}$$

Now, we can make the following replacements: $S^\dagger = \mathbb{1} - iT^\dagger$ and $S \mapsto iT$. The asymmetry between S and $S^\dagger$ has a simple interpretation: $R_3(j_3 j_2)$ requires particle 3 to be in the future lightcone of particle 2; thus particles 1 and 2 must interact together *before* particle 3 is emitted, since all particles are stable and cannot freely radiate. Putting everything together, dropping the forward terms, and inserting a

complete basis of asymptotic states X, we have

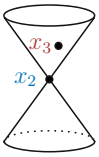
$$
\begin{aligned}
-\langle 0|R_5(j_5 j_4)R_3(j_3 j_2)a_1^\dagger|0\rangle &= i(-i)\langle 0|a_5 a_4 T^\dagger a_3 T a_2^\dagger a_1^\dagger|0\rangle \\
&= \sum_X \!\!\!\!\!\!\int \langle 0|a_5 a_4 T^\dagger|X\rangle\langle X|a_3 T a_2^\dagger a_1^\dagger|0\rangle \\
&= \quad 4 \quad 5 \quad -i\mathcal{M}^\dagger \quad X \quad i\mathcal{M} \quad 2 \quad 1 \quad 3 \\
&\equiv \mathrm{Cut}_{45}\boldsymbol{\delta}_5 \,.
\end{aligned}
\tag{5.2.51}
$$

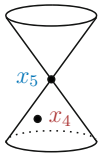
Thus, the correlator (5.2.49) is just a particular way of writing a unitarity cut. It turns out that this representation has a very interesting feature. To see it, let us examine a bit its Fourier transform

$$
\int \mathrm{d}^{\mathrm{D}}x_5 \mathrm{d}^{\mathrm{D}}x_4 \mathrm{d}^{\mathrm{D}}x_3 \mathrm{d}^{\mathrm{D}}x_2 \mathrm{e}^{-i(p_5\cdot x_5 + p_4\cdot x_4 + p_3\cdot x_3 + p_2\cdot x_2)}\langle 0|R_5(j_5 j_4)R_3(j_3 j_2)a_1^\dagger|0\rangle \,, \tag{5.2.52}
$$

and perform, say, the constant timelike shifts $x_3 \mapsto x_3 + \xi$ and $x_4 \mapsto x_4 + \xi$. Now, physically, if particles 3 and 4 are measured by a detector asymptotically far in the future ($\xi \to \infty$), we expect some s_{34}-channel non-analyticity in the momentum space representation (5.2.52).

However, because of the explicit dependence of the integrand on $R_5(j_5 j_4)$ and $R_3(j_3 j_2)$, the ξ-range is constrained as follows:

$R_3(j_3 j_2) \rightsquigarrow$ 

and $R_5(j_5 j_4) \rightsquigarrow$ 

(5.2.53)

In other words, the integrand has support only for

$$
\infty \geqslant x_3^\pm - x_2^\pm > 0 \mapsto \xi^\pm > x_2^\pm - x_3^\pm \geqslant -\infty \,, \tag{5.2.54a}
$$

$$
\infty \geqslant x_5^\pm - x_4^\pm > 0 \mapsto \infty \geqslant x_5^\pm - x_4^\pm > \xi^\pm \,, \tag{5.2.54b}
$$

such that

$$
\infty \geqslant x_5^\pm - x_4^\pm > \xi^\pm > x_2^\pm - x_3^\pm \geqslant -\infty \,. \tag{5.2.55}
$$

Furthermore, since ξ is timelike, we have a constraint on $\boldsymbol{\xi}_\perp$

$$
\xi^2 = -\xi^+\xi^- + \boldsymbol{\xi}_\perp^2 < 0 \implies -\sqrt{\xi^+\xi^-} < |\boldsymbol{\xi}_\perp| < \sqrt{\xi^+\xi^-} \,. \tag{5.2.56}
$$

Conditions (5.2.55) and (5.2.56) put together prohibit past and future infinite timelike translations (i.e., limits $|\xi| \to \infty$). This indicates that we should not expect any s_{34}-channel singularity. In other words, we just argued that

$$\text{Disc}_{s_{34}} \text{Cut}_{45} = 0\,. \tag{5.2.57}$$

This is an example of the so-called *Steinmann relations*. Of course, going through a similar reduction argument on the correlator $\langle 0|R_4(j_5 j_4)R_3(j_3 j_2)a_1^\dagger|0\rangle$, one could similarly argue that

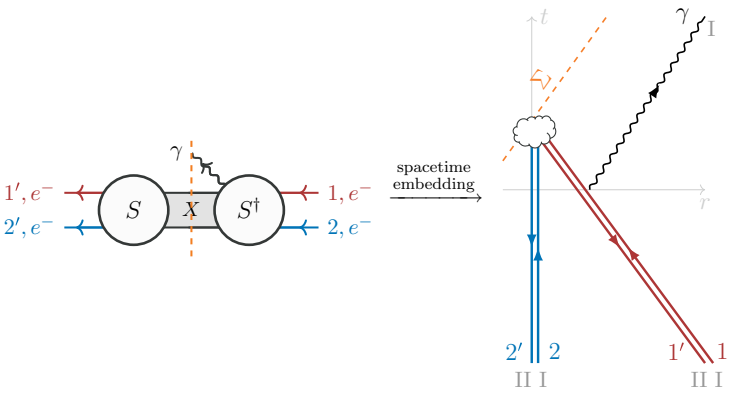
$$\text{Disc}_{s_{35}} \text{Cut}_{45} = 0\,. \tag{5.2.58}$$

As before, this discussion suggests that the existence of multiple reduction formulae is central in revealing the full analytic properties of amplitudes.

In the last few examples we examined, the Schwinger–Keldysh framework was not used. To demonstrate how it can be used to extract quantitative details about OTOCs, we next apply it to analyze the infrared (IR) divergences of the five-point in-in observable in (5.1.23).

5.2.3 Infrared Divergences of In-In QED Observables from the Timefold

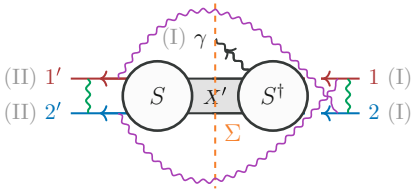
The purpose of this section is to analyze how the Schwinger–Keldysh picture introduced above can be used to understand the infrared (IR) properties of out-of-time-ordered observables. To do so, we will focus on a particular in-in QED observable, namely the inclusive measurement ("waveform") of a photon in the background of two electrically charged particles (which we will take to be electrons); the analogous calculation in gravity is discussed in [2] (see also [19, Sec. 4]). Using the blob pictures introduced earlier, this observable is represented as



(5.2.59)

Here, the spacelike surface Σ is arbitrary and need only to be in the causal past of photon measurement; in the Schwinger–Keldysh picture (right panel above), it plays the role of the "turning point" from fold I to II.

Our goal now is to extract the IR divergence of this observable from the Schwinger–Keldysh formalism. However, before we go into the quantitative details, let us spell out the physical expectation. Physically, we expect the IR divergences to arise from *asymptotic* regions where the electrons follow eikonal lines and interact weakly via soft-photon exchanges. More precisely, these regions are associated to the $2 + 2 = 4$ Feynman diagrams that describe the long-range photon exchanged between the electrons on the same timefold *and* those on different timefolds



(5.2.60)

where X' is the complete set of states, apart from the photons that cross the cut at Σ.

Note that self-energy diagrams, which account for particles moving in their own fields, are not included. This will be explained in more detail later on. Moreover, soft interactions involving the X' states do not need to be considered since these cancel out upon inclusively summing over X'—note that X' never even appears in the Schwinger–Keldysh calculation if we place the timefold Σ early enough—this is the essence of the celebrated Bloch–Nordsieck theorem [20].

Now, a canonical way to extract the IR divergence is therefore to compute these one-loop Feynman diagrams, sum them up, and use QED exponentiation identities to get the all-order result. We can check that this procedure leads to the same result as the one we are about to derive by considering Wilson lines in the Schwinger–Keldysh formalism (see (5.2.73) below).

As hinted above, what we need to compute is the vacuum expectation value of the electrically charged eikonal lines on the Schwinger–Keldysh contour $\mathcal{C}$:

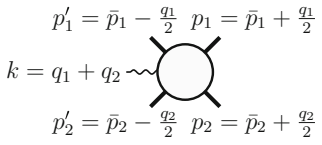
$$\begin{aligned}\mathrm{Exp}^{\mathcal{C}} &= \langle 0|\,\mathcal{C}\exp\left[\frac{ie}{\hbar}\sum_{j\in\{1,2,1',2'\}}\int \mathrm{d}\tau p_j^\mu A_\mu(\tau p_j)\right]|0\rangle \\ &= \langle 0|\,\mathcal{C}\exp\left[\frac{ie}{\hbar}\sum_{j\in\{1,2\}}\int_{-\infty}^{\Sigma}\mathrm{d}\tau(p_j^\mu A_\mu(\tau p_j) - p_{j'}^\mu A_\mu(\tau p_{j'}))\right]|0\rangle\ .\end{aligned} \tag{5.2.61}$$

Above, e and p_j^μ are the electrons' charge and momenta (respectively), A_μ denotes the gauge field of the soft radiation, while τ parameterizes the electrons' straight line trajectories ($x_j^\mu = \tau p_j^\mu$). We will see later how the four Feynman diagrams

mentioned above (see (5.2.60)) arise in the leading-order term in the eikonal regime (which we will refer to below as the "*classical part*") in the Schwinger–Keldysh picture.

We will focus here on the "classical regime" where $p_i - p_i' = \mathcal{O}(\hbar)$ (physically related to the width of wavepackets in (5.1.21)); this approximation could be relaxed if desired.

Thus, we first extract the classical part of the integrand in (5.2.61). To do so, it is convenient to introduce the following kinematics



(5.2.62)

In the classical regime, the momentum transfer is small, meaning that $q_i = \mathcal{O}(\hbar)$. Similarly, the difference in gauge fields is correspondingly small

$$A_\mu^{\text{diff}}(x) = A_\mu^{\text{I}}(x) - A_\mu^{\text{II}}(x) = \mathcal{O}(\hbar)\,. \tag{5.2.63}$$

Note that by momentum conservation, the barred momenta are orthogonal to the respective momentum transfer: $q_j \cdot \bar{p}_j = 0$.

Using the above results, an elementary Taylor expansion around $q_j^\mu = 0$ gives

$$p_j^\mu A_\mu(\tau p_j) - p_{j'}^\mu A_\mu(\tau p_{j'}) = \underbrace{\bar{p}_j^\mu A_\mu^{\text{diff}}(\tau \bar{p}_j) + q_j \cdot \frac{\partial}{\partial \bar{p}_j} \bar{p}_j^\mu A_\mu^{\text{avg}}(\tau \bar{p}_j)}_{\mathcal{O}(\hbar)} + \mathcal{O}(\hbar^2)\,, \tag{5.2.64}$$

where we introduce the average of fields $A_\mu^{\text{avg}}(x) = \frac{1}{2}(A_\mu^{\text{I}}(x) + A_\mu^{\text{II}}(x))$. Plugging this back into (5.2.61), we have (up to effects suppressed in $\hbar$)

$$\text{Exp}^{\mathcal{C}} = \langle 0|\, \mathcal{C} \exp\left[\frac{ie}{\hbar} \sum_{j\in\{1,2\}} \int_{-\infty}^{\Sigma} \mathrm{d}\tau \left(\bar{p}_j^\mu A_\mu^{\text{diff}}(\tau \bar{p}_j) + q_j \cdot \frac{\partial}{\partial \bar{p}_j} \bar{p}_j^\mu A_\mu^{\text{avg}}(\tau \bar{p}_j)\right)\right] |0\rangle\,. \tag{5.2.65}$$

Now, since both terms in the integrand are of $\mathcal{O}(\hbar)$, it is clear that the exponential argument is purely classical. Consequently, from now on we set $\hbar = 1$.

From there we proceed by making the following observation: in the IR, the gauge fields are Gaußian and thus *fully* characterized by their two-point functions. This

means that the average of the exponential is equal to the exponential of half the "variance"

$$\mathrm{Exp}^{\mathcal{C}} = \exp\Big[\frac{-e^2}{2}\sum_{j\in\{1,2\}}\int_{-\infty}^{\Sigma}\mathrm{d}\tau\mathrm{d}\tau'\,\langle 0|\,\mathcal{C}\Lambda_j(\tau)\Lambda_j(\tau')\,|0\rangle\Big]. \tag{5.2.66}$$

where we introduce the shorthand $\Lambda_j(\tau) = \bar{p}_j^{\mu} A_{\mu}^{\mathrm{diff}}(\tau\bar{p}_j) + q_j\cdot\frac{\partial}{\partial\bar{p}_j}\bar{p}_j^{\mu}A_{\mu}^{\mathrm{avg}}(\tau\bar{p}_j)$. This is the usual mechanism underlying the exponentiation of infrared divergences in QED and here we are simply applying it to the waveform.

Next, expanding $\langle 0|\,\mathcal{C}\Lambda_j(\tau)\Lambda_j(\tau')\,|0\rangle$ leads to 16 terms, many of which (14) vanish. For some terms, the vanishing is mathematically obvious. For example, from the largest time equation in (5.2.24), we have

$$\langle 0|\,\mathcal{C}A_{\mu}^{\mathrm{diff}}(x)A_{\mu}^{\mathrm{diff}}(y)\,|0\rangle = 0\,, \tag{5.2.67}$$

while the $A^{\mathrm{avg}}A^{\mathrm{avg}}$ terms come with two powers of q which makes them sub-classical compared to the terms to be considered shortly. For the remaining $A^{\mathrm{diff}}A^{\mathrm{avg}}$ terms, a calculation is needed. Two of them involve "fields moving in their own fields" which vanish in dimensional regularization as they give scaleless integrals, as will be made explicit shortly. Thus the calculation reduces to

$$\mathrm{Exp}^{\mathcal{C}} = \exp\Big[-e^2\sum_{j\in\{1,2\}} q_j\cdot\frac{\partial}{\partial\bar{p}_j}\bar{p}_j^{\mu}\bar{p}_{\bar{j}}^{\nu}\int_{-\infty}^{\Sigma}\mathrm{d}\tau\mathrm{d}\tau'\,\langle 0|\,\mathcal{C}A_{\mu}^{\mathrm{avg}}(\tau\bar{p}_j)A_{\nu}^{\mathrm{diff}}(\tau'\bar{p}_{\bar{j}})\,|0\rangle\Big], \tag{5.2.68}$$

where $\bar{j}\in\{1,2\}\setminus j$ and the factor of $\frac{1}{2}$ is cancelled by the factors of 2 in the surviving cross-terms. Let us point out that expanding the bracket in the I–II basis would return a total of $2+2=4$ terms, which are in one-to-one correspondence with the four physically relevant Feynman diagrams anticipated earlier in (5.2.60). Using retarded propagators reduces the number of diagrams that need to be considered in the classical limit.

Next, in order to perform the τ integrals in (5.2.68), we need to evaluate the leftover retarded propagators (defined earlier in (5.2.21)) in, e.g., the Feynman gauge

$$\langle 0|\,\mathcal{C}A_{\mu}^{\mathrm{avg}}(x)A_{\nu}^{\mathrm{diff}}(y)\,|0\rangle = \eta_{\mu\nu}\int\frac{\mathrm{d}^{\mathrm{D}}p}{(2\pi)^{\mathrm{D}}}\frac{-i\mathrm{e}^{-ip\cdot r}}{p^2-i\varepsilon p^0}\qquad\text{with } r=x-y\,. \tag{5.2.69}$$

A useful observation to perform this integral is to note that the result is Lorentz invariant. Assuming that r is timelike, we can therefore boost in a frame where $\mathbf{r}=0$, perform the integral in p^0 using the residue theorem, use spherical coordinates to perform the remaining spatial integrations and then boost the answer back to $\mathbf{r}\neq 0$.

Setting $\mathrm{D} = 4$ at the very end of this procedure gives

$$\langle 0|\, \mathcal{C} A_\mu^{\mathrm{avg}}(x) A_\nu^{\mathrm{diff}}(y)\, |0\rangle = \frac{\eta_{\mu\nu}\theta_{xy}}{2\pi i}\delta[(x-y)^2]\,. \tag{5.2.70}$$

Now using this expression, the (divergent) τ integrals can be performed, leading to

$$\mathrm{Exp}^{\mathcal{C}} = \exp\Big[\frac{-ie^2}{4\pi}\sum_{j\in\{1,2\}} q_j\cdot\frac{\partial}{\partial\bar{p}_j}\frac{\bar{y}}{\sqrt{\bar{y}^2-1}}\int_{-\infty}^{\Sigma}\mathrm{d}\log\tau\Big]\,, \tag{5.2.71}$$

where we have used $\bar{p}_i = \bar{m}_i\bar{v}_i$ as well as $\bar{y} = -\bar{v}_1\cdot\bar{v}_2$. To regulate the logarithmically divergent integral, we introduce a cutoff by replacing the lower bound with the spatial separation between the incoming states ($\Delta x_{\mathrm{in}}^\mu = \tau_{\min}\Delta p^\mu$) at the moment when they are shot into the bulk, while the upper bound is replaced by the impact parameter ($b^\mu = \tau_{\max}\Delta p^\mu$). This leads to

$$\mathrm{Exp}^{\mathcal{C}} = \exp\Big[\frac{-ie^2}{4\pi}\sum_{j\in\{1,2\}} q_j\cdot\frac{\partial}{\partial\bar{p}_j}\frac{\bar{y}}{\sqrt{\bar{y}^2-1}}\log\frac{\tau_{\max}}{\tau_{\min}}\Big]\,. \tag{5.2.72}$$

Finally, computing the derivatives yields the logarithmically divergent time shift

$$\mathrm{Exp}^{\mathcal{C}} = \exp\Big[-i\frac{e^2}{4\pi}\frac{k\cdot(\bar{p}_1+\bar{p}_2)}{\bar{m}_1\bar{m}_2}\frac{1}{(\bar{y}^2-1)^{3/2}}\log\frac{\tau_{\max}}{\tau_{\min}}\Big] \equiv \exp\Big[-i\Delta t\,\Delta E\Big]\,, \tag{5.2.73}$$

where in the rest frame $\Delta E = k^0$. We see that if we shot the initial states from infinitely far in the past at fixed energy, the in-in observable experiences a large logarithmic time delay (for a positron-electron pair, it would be a time advance). This is due to the repulsive $\propto e^2/r^2$ between them, which delays the moment of collision relative to the straight-line trajectories of non-interacting particles which would be sent with identical initial conditions. We point out that the conventional (time-ordered) amplitude $\mathcal{M}_{e^-e^-\gamma\leftarrow e^-e^-}$ does not experience this time delay; this is because the phase in (5.2.73) is exactly canceled by that from interactions between the out states (which comes with an opposite sign by momentum conservation). Instead, the time-ordered amplitude would exhibit a large *real* exponent $\mathrm{e}^{-\frac{1}{2}N_\gamma}$ that represents the probability of *not* emitting any photon. We see that infrared divergences for the inclusive waveform and exclusive 5-point amplitude describe vastly different physics. A similar setup in gravity is discussed in the contribution [19, Sec. 4].

5.3 The FBI Local Transform and Analyticity of Multi-Point Scattering Amplitudes

Simon Caron-Huot

The purpose of this final lecture is to provide a rough overview of [10, 21] and to discuss the implications of causality on the analyticity of $m \leftarrow n$ scattering amplitudes.

One of the main historical hurdles in this program is the fact that conventional Fourier transforms of position-space correlators do not show clear convergence on the mass shell (this was pointed out earlier in Sect. 5.2.2). The challenge, therefore, involves devising a variant of the traditional Fourier transform that not only converges on the mass shell but also possesses convergence properties robust enough to study analyticity deep in the complex kinematic space.

To date, a clear and practical understanding of how to realize such a construction remains blurry. However, we believe that a key tool in achieving this goal is the so-called *Fourier–Bros–Iagolnitzer* (FBI) transform, or *local Fourier transform* [21].

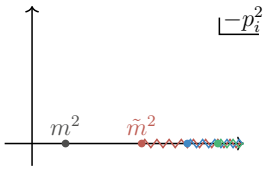
Below, we sketch a (very!) brief yet comprehensive overview of this intricate topic, which we believe could be useful for anyone targeting this goal. See also [3, App. A].

5.3.1 Why (and When) Are Amplitudes Analytic?

The story begins with the conventional Fourier transform of the (initially off-shell) position-space correlator:

$$G(p_1, \ldots, p_n) = \int \langle 0|T(j_1 \ldots j_n)|0\rangle \prod_{i=1}^{n} \mathrm{d}^{\mathrm{D}} x_i \mathrm{e}^{-ip_i \cdot x_i} \, .n \tag{5.3.1}$$

We assume that the first singularity of the two-point functions of currents j_i is at some mass $\tilde{m} > m$ which is strictly separated from the mass shell, i.e., that the two-point function of unamputated fields looks like:



(5.3.2)

To streamline the discussion, we first set $n = 4$ and focus on $34 \leftarrow 12$ kinematics.

Example 5.3.1 (Four-Point) Under this assumption, stability ensures, for example

$$\mathcal{G}_1 = \langle 0|T(j_2 j_3 j_4)T(j_1)|0\rangle = \sum\!\!\!\!\!\!\int_X \langle 0|\, T(j_2 j_3 j_4)\, |X\rangle\, \langle X|\, T(j_1)\, |0\rangle = 0\,. \tag{5.3.3}$$

Therefore, near the mass shell (5.3.1) can also be written as

$$G(p_1 p_2 p_3 p_4) = \int (\langle 0|T(j_1 j_2 j_3 j_4)|0\rangle - \mathcal{G}_1) \prod_{i=1}^{4} \mathrm{d}^{\mathrm{D}} x_i\, \mathrm{e}^{-ip_i\cdot x_i} \equiv G_1\,. \tag{5.3.4}$$

The interesting thing about (5.3.4) is that its support is smaller because the right-hand side manifestly vanishes unless 1 is in the future of either 2, 3, 4:

$$\mathrm{supp}(G_1) = \{x : x_1 \succcurlyeq x_2 \vee x_3 \vee x_4\}\,. \tag{5.3.5}$$

Now, we could write (5.3.1) in various other ways that have different threshold properties, by subtracting different products that vanish below the threshold. There are in total eight terms similar to (5.3.3), namely

$$\mathcal{G}_{1\leqslant i\leqslant 4} = \langle 0|T(\prod_{\substack{k=1\\k\neq i}}^{4} j_k)T(j_i)|0\rangle \quad \text{and} \quad \mathcal{G}_{5\leqslant i\leqslant 8} = \langle 0|T(j_i)T(\prod_{\substack{k=1\\k\neq i-4}}^{4} j_k)|0\rangle\,. \tag{5.3.6}$$

In $34 \leftarrow 12$ kinematics, we further have

$$\mathcal{G}_9 = \langle 0|T(j_1 j_3)T(j_2 j_4)|0\rangle = 0\,, \tag{5.3.7a}$$
$$\mathcal{G}_{10} = \langle 0|T(j_2 j_4)T(j_1 j_3)|0\rangle = 0\,, \tag{5.3.7b}$$
$$\mathcal{G}_{11} = \langle 0|T(j_1 j_4)T(j_2 j_3)|0\rangle = 0\,, \tag{5.3.7c}$$
$$\mathcal{G}_{12} = \langle 0|T(j_2 j_3)T(j_1 j_4)|0\rangle = 0\,, \tag{5.3.7d}$$
$$\mathcal{G}_{13} = \langle 0|T(j_1 j_2)T(j_3 j_4)|0\rangle = 0\,. \tag{5.3.7e}$$

Note that $\mathcal{G}_{14} = \langle 0|T(j_3 j_4)T(j_1 j_2)|0\rangle \neq 0$ in these kinematics. Let us emphasize once more that

$$G_i = \int (\langle 0|T(j_1 j_2 j_3 j_4)|0\rangle - \mathcal{G}_i) \prod_{i=1}^{4} \mathrm{d}^{\mathrm{D}} x_i \mathrm{e}^{-ip_i\cdot x_i} \tag{5.3.8}$$

all define different functions and they only agree for real momenta which happen to be below certain thresholds.

What comes next relies on picking a point p in $34 \leftarrow 12$ kinematics. Define the *essential support* near p as the intersection of the support of all representations which agree near p:

$$\mathrm{ES}_p = \bigcap_{i=1}^{13} \mathrm{supp}(G_i(x))\,. \tag{5.3.9}$$

The amazing claim is that the analyticity properties of G in a complex neighborhood of p are the same as those of a Fourier transform of a function supported in ES_p.

The calculation of such intersections is an interesting exercise in Boolean mathematics (try it!). Some G_i force certain points to be in the future of others, and some other G_i's impose the opposite constraint. The result is that we need: $x_1 = x_2$; $x_3 = x_4$; and a certain causality relation between these two points:

$$\mathrm{ES}_{34\leftarrow 12} = \{x : x_4 \asymp x_3 \not\succcurlyeq x_1 \asymp x_2\} \qquad \begin{matrix} 3 & & & & 2 \\ & \searrow & & \swarrow & \\ & & \bullet & & \\ & \nearrow & & \searrow & \\ 4 & & & & 1 \end{matrix} \tag{5.3.10}$$

The physical interpretation of (5.3.10) is that near any point p in $34 \leftarrow 12$ kinematics, $G(p_1p_2p_3p_4)$ has the same analyticity properties as the Fourier transform of the shown tree-level s-channel Feynman diagram, in which the internal state is allowed to propagate in any direction inside the light cone.

To see the implied analyticity properties, write $y \asymp x_3 \asymp x_4$ and $x \asymp x_1 \asymp x_2$ such that

$$G(p_1p_2p_3p_4) \simeq \int_{y\not\succcurlyeq x} \mathrm{d}^{\mathrm{D}}x\, \mathrm{d}^{\mathrm{D}}y\, \mathrm{e}^{-i(p_3+p_4)\cdot(y-x)}(\ldots)\,. \tag{5.3.11}$$

The integral converges exponentially provided $\mathrm{Re}(-i(p_3+p_3)\cdot(y-x)) < 0$ throughout, which is guaranteed if $\mathrm{Im}(p_3+p_3)^\mu \in V^+$. In $34 \leftarrow 12$ kinematics, this only amounts to a constraint on s:

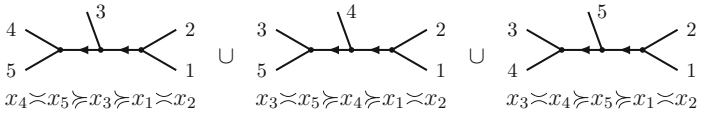
$$\begin{aligned} \mathrm{Im}(s) &= \mathrm{Im}(-(p_3+p_4)^2) \\ &= -2\underbrace{\mathrm{Im}(p_3+p_4)}_{\in V^+}\cdot\underbrace{\mathrm{Re}(p_3+p_4)}_{\in V^+} > 0\,. \end{aligned} \tag{5.3.12}$$

Since the amplitude is a function of Mandelstam invariants, this means that (5.3.11) converges, provided that we have $s + i\varepsilon$, but we can have either $t \pm i\varepsilon$. This is a statement about a small complex neighborhood of an arbitrary real point p.

Note that the integrand $(\cdots)$ in (5.3.11) is not quite the same as the original correlator, however it will be constructed explicitly below as part of the proof.

Of course, a similar story persists at higher multiplicities. By automating the intersection calculus, we find the following essential supports:

Example 5.3.2 (Five-Point) Repeating this exercise for $n = 5$, we find that the essential support of $\langle 345|S|12|345|S|12\rangle$ is contained within

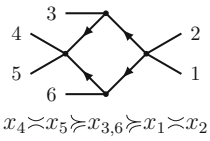


(5.3.13)

Each contribution is analytic in some complex cone around real momenta. When the intersection of these three cones and the mass shell is nonempty, the amplitude is analytic; otherwise, it is a priori just the sum of three functions which are each analytic in different cones; this example is further discussed in [10].

We notice that once again the essential support diagrams are tree-level diagrams. The situation changes at six-point.

Example 5.3.3 (Six-Point) Repeating this exercise for $n = 6$ is a more challenging combinatorial exercise. We find that the essential support of $\langle 3456|S|12|3456|S|12\rangle$ contains loop-like diagrams, e.g.,



(5.3.14)

The physical significance of such diagrams remains unclear.

5.3.2 The FBI Transform

How can we obtain Fourier representations like (5.3.11) that converge on the mass shell? The FBI (Fourier–Bros–Iagolnitzer) transform gives a nice constructive procedure [21] that is essentially a judicious application of Stoke's theorem. The procedure has been used by [10] to prove analyticity in the vicinity of a real point p but we suspect that it may also be a good tool to study physics deep into the complex plane.

The construction uses the following ingredients. Let

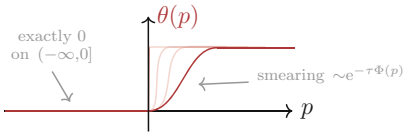
$$\begin{aligned} R &= \text{an open set in real momentum space}\,, \\ \Phi(p) &= \text{a localizing function such that } \mathrm{Re}(\Phi(p)) \begin{cases} < 1 \text{ inside } R = \mathring{R}\,, \\ = 1 \text{ on } \partial R\,, \end{cases} \\ \{G_i(p)\} &= \text{a set of functions that agree on } R \text{ (see, e.g., (5.3.8))}. \end{aligned} \tag{5.3.15}$$

Denote ES_R the intersection of the supports of all $G_i(x)$. Then, for any $\tau > 0$, the FBI transform is defined by:

$$\mathcal{F}_{\Phi,\tau}(x) = \int \frac{\mathrm{d}^{\mathrm{D}}p}{(2\pi)^{\mathrm{D}}} \mathrm{e}^{ip\cdot x - \tau\Phi(p)} G_i(p)\theta(p \in R)\,. \tag{5.3.16}$$

There are a few important comments to make about this definition:

1. The transform $\mathcal{F}_{\Phi,\tau}(x)$ does not depend on the index i; this is because of the explicit step function, which is only non-zero on the region R where all the G_i's agree. For instance, for the four-point example discussed in Sect. 5.3.1, we would take $\theta(p \in R) = \theta(p_1^2 + \tilde{m}_1^2)\theta(p_2^2 + \tilde{m}_2^2)\ldots$
2. Naively, we might think that $\mathcal{F}_{\Phi,\tau=0}(x)$ enjoys the support properties of each $G_i(x)$; that is, that $\mathcal{F}_{\Phi,\tau=0}(x) = 0$ for $x \notin \mathrm{ES}_R$. The only obstruction to this is that $\theta(p \in R)$ is not an analytic function near the boundary of R: this prevents $\mathcal{F}_{\Phi,\tau}(x)$ from decaying exponentially in any direction.
3. Technically, it is straightforward to replace $\theta(p \in R)$ by a smooth approximation to a step function, so as to avoid ill-defined products of distributions. However, no amount of smoothness will grant us exponential decay in directions outside ES_R. The key will be to use τ. Only by increasing τ proportionally to the distance from ES_R can we suppress the contributions from the sharp transition of the step function at the boundary ∂R:



(5.3.17)

   The boundary contributions are then manifestly suppressed by

$$\mathrm{e}^{-\tau\Phi(p)}|_{p\in\partial R} = \mathrm{e}^{-\tau}\,. \tag{5.3.18}$$

4. This will turn into get exponential decay outside of ES_R by choosing suitable $\tau(x) \propto |x|$.

Physically, for Gaußian $\Phi(p)$, $\tau(x) \propto |x|$ means roughly that we convolve the correlator against wavepackets whose width in x-space grows like $\sqrt{|x|}$. This exponentially suppresses unwanted contributions from ∂R because these oscillate at the "wrong frequency" (compared with frequencies $p \in R$ which we are interested in).

The second step in the construction is to find an inverse transform for a non-constant $\tau(x)$.

Thanks to standard Fourier inversion theorems, $G(p)$ for $p \in R$ can be recovered by integrating (5.3.16) along any constant-τ slice, and then multiplying the result

by $e^{\tau\Phi(p)}$:

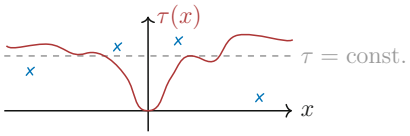
$$G(p) = \int d^D x\, e^{-ip\cdot x+\tau\Phi(p)}\mathcal{F}_{\Phi,\tau}(x) \qquad (\tau\text{=constant}). \tag{5.3.19}$$

This does not yet grant us any analyticity in p, because $\mathcal{F}_{\Phi,\tau}(x)$ decays with τ but not with x.

Remarkably, the inverse transform (5.3.19) can now be deformed to a general surface $\tau(x)$. The construction is surprisingly simple and amounts to turning the integrand into the pullback of a closed n-form on $(\tau, x) \in \mathbb{R}^{D+1}$ space:

$$G(p) = \underbrace{\int_{\tau=\tau(x)} e^{-ip\cdot x+\tau(x)\Phi(p)}\left(\mathcal{F}_{\Phi,\tau(x)}(x)d^D x + \sum_{i=1}^{D}\mathcal{F}^i_{\Phi,\tau(x)}(x)d\tau d^{D-1}x_i\right)}_{\text{(any admissible } \tau(x))}. \tag{5.3.20}$$

We invite the reader to look at [21] for details on the construction of $\mathcal{F}^i_{\Phi,\tau(x)}(x)$ ensuring that the parenthesis is a *closed* form. By Stokes' theorem, the integral is then independent of the surface $\tau(x)$

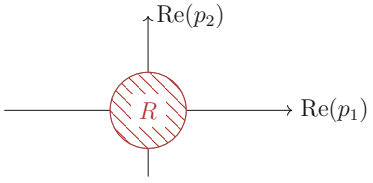


(5.3.21)

and therefore equal to $G(p)$ by the above comments, provided that the surface is homologous to a constant-τ slice.

This is a practical result because it implies that we have the freedom to select the *optimal* $\tau(x)$—i.e., that which produces the strongest exponential decay with x—to make the analyticity properties of $G(p)$ manifest in possibly large (complex) domains in p. Below, we will illustrate how the choice of $\Phi(p)$ can greatly influence the size of the analyticity domains.

Example 5.3.4 (Gaußian Φ and Edge-Of-The-Wedge Theorem) As a simple application, to be continued below, consider two functions $G_1(p_1, p_2)$ and $G_2(p_1, p_2)$ of two complex variables ($G_i : \mathbb{C}^2 \to \mathbb{C}$), which coincide in a real region R which we take to be the unit ball:



(5.3.22)

For R the unit ball, we can choose Φ to be a Gaußian $\Phi(p_1, p_2) = p_1^2 + p_2^2$, such that $\Phi(\partial R) = 1$ and $0 \leqslant \Phi(\overset{\circ}{R}) < 1$. The FBI transform and its inverse read explicitly

$$\mathcal{F}_\tau(x_1, x_2) = \int_{p_1^2+p_2^2\leqslant 1} \frac{\mathrm{d}p_1 \mathrm{d}p_2}{(2\pi)^2} \mathrm{e}^{ip_1x_1+ip_2x_2-(p_1^2+p_2^2)\tau} G_i(p_1, p_2), \tag{5.3.23a}$$

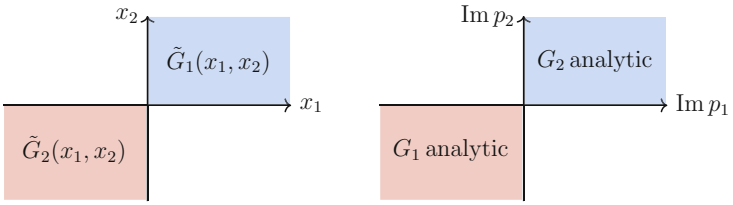
$$G(p_1, p_2) = \int \mathrm{d}x_1 \mathrm{d}x_2 \, \mathrm{e}^{-ip_1x_1-ip_2x_2+(p_1^2+p_2^2)\tau} \mathcal{F}^*(x_1, x_2)\,, \tag{5.3.23b}$$

where

$$\mathcal{F}^*(x_1, x_2) = \left(1 + \sum_{j=1}^{2} \frac{\mathrm{d}\tau}{\mathrm{d}x_j} \left(ip_j - \frac{\partial}{\partial x_j}\right)\right) \mathcal{F}_\tau(x_1, x_2)\Big|_{\tau=\tau(x_1,x_2)}. \tag{5.3.24}$$

The $\mathrm{d}\tau/\mathrm{d}x_j$ term originates from the correction term in (5.3.20) and ensures invariance under small deformation of the surface $\tau(x_1, x_2)$. (One should set $\tau \mapsto \tau(x)$ in $\mathcal{F}$ only after taking the partial derivative.)

We will now use the FBI transform and its inverse to give a proof of the celebrated *edge-of-the-wedge* theorem. For this we will suppose, moreover, that the conventional Fourier transforms of $G_1(p_1, p_2)$ and $G_2(p_1, p_2)$ are supported, respectively, on the first and third quadrants in x-space (and thus analytic in corresponding quadrants of Im p):



(5.3.25)

The basic *edge-of-the-wedge* question is: what analyticity near R does the agreement between $G_1(p_1, p_2)$ and $G_2(p_1, p_2)$ for real p grants us?

Since the intersection of the x-supports is a single point $\mathrm{ES}_R = \{0\}$, the logic of the essential support suggests that the function should enjoy near R the analytic properties of a Fourier integral supported on a point, i.e., be analytic in a full complex neighborhood of R. This is a non-trivial extension of the analyticity domain in the right panel of (5.3.25), where we add some small disk around the origin. This is exactly the content of the *edge-of-the-wedge theorem*, which we will now explain from the FBI perspective.

The key question to ask is: how small can we make $\mathrm{e}^{-\tau(x_1,x_2)}$ outside the first quadrant, if we take the transform (5.3.23a) of G_1? And how small can we make

$e^{-\tau(x_1,x_2)}$ outside the third quadrant if we use the transform with G_2? Since the two transforms agree, our estimate of $\mathcal{F}_\tau$ will be the best of both.

Consider the transform starting with $G_1(p_1, p_2)$ in (5.3.23a) and suppose $x_1 < 0$. Because G_1 is analytic in the third quadrant of Im p, the factor $e^{ip_1x_1}$ can be made smaller by deforming the contour into the lower-half p_1-plane, where the integrand is analytic. As usual with Fourier integrals, we estimate that it is "as small as we can make it by deforming the contour."

Since the contour is pinned to the real boundary of ∂R where $\Phi(\partial R) = 1$, the *optimal* decay we can hope to achieve is $e^{-\tau}$. Thus, the useful complex integration contours are contours of uniform suppression:

$$\mathrm{Re}(-ip\cdot x + \tau\Phi(p)) = \tau\,. \tag{5.3.26}$$

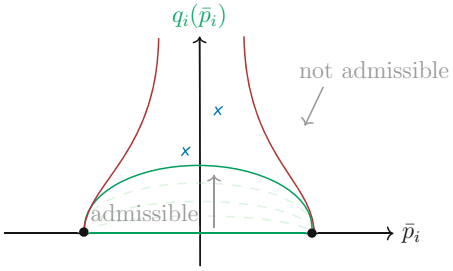
Following [21], we restrict attention to contours where the imaginary part of p is along a single direction q, such that $p_j = \bar{p}_j + i\tau\lambda(\bar{p}_j)q_j$, where $\bar{p}_j$, q_j and λ are all real (and we always keep x's and τ real). The relation (5.3.26) becomes

$$\lambda(\bar{p})\frac{q\cdot x}{\tau} + \mathrm{Re}\,\Phi(\bar{p} + iq\lambda(\bar{p})) = 1\,. \tag{5.3.27}$$

We can eliminate $\frac{q\cdot x}{\tau}$ by rescaling λ and setting $\tilde{q} = q\frac{\tau}{q\cdot x}$, to characterize the constant-suppression contours as:

$$\Gamma_{\Phi,\tilde{q}} = \{p = \bar{p} + i\lambda(\bar{p})\tilde{q} :\ \bar{p} \in R \text{ and } \lambda(\bar{p}) = 1 - \mathrm{Re}\,\Phi(\bar{p} + i\tilde{q}\lambda(\bar{p}))\}\,. \tag{5.3.28}$$

We say that $\Gamma_{\Phi,\tilde{q}}$ is *admissible* if it is a sensible contour (i.e., λ real and positive everywhere) and is a deformation of the original one with $\lambda = 0$ in (5.3.23a). Contours with small enough $\tilde{q}$ are, of course, always admissible. With increasing magnitude of $\tilde{q}$ we may need to stop at a singularity of $G_i(p)$ or if λ becomes complex. This is summarized in the following cartoon:



(5.3.29)

Finally, for a given x, we find the optimal decay $e^{-\tau}$ by scanning over admissible $\tilde{q}$:

$$\tau_{\max}(x) = \max_{\tilde{q}\ \text{admissible}} \tilde{q}\cdot x\,. \tag{5.3.30}$$

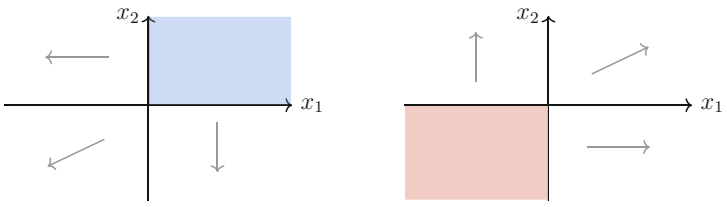
In the case of Gaußian Φ, the contours in (5.3.28) can be found analytically and are sensible provided that $\sqrt{\tilde{q}_1^2 + \tilde{q}_2^2} \leqslant \frac{1}{2}$. In order to be admissible for the G_1 version of (5.3.23a), the vector q must also lie within the third quadrant; for the G_2 version, it must lie in the first quadrant. Since the two versions agree, the set of admissible contours is the union:

$$\Gamma_{\Phi,\tilde{q}} \text{ is admissible } \Leftrightarrow \sqrt{\tilde{q}_1^2 + \tilde{q}_2^2} < \frac{1}{2} \text{ and } \tilde{q}_1\tilde{q}_2 \geqslant 0\,. \tag{5.3.31}$$

The optimal decay in (5.3.30) is then $\mathrm{e}^{-\tau_{\max}(x)} = \mathrm{e}^{-C|x|}$ where C is an $\mathcal{O}(1)$ constant depending on the direction of x. Crucially, $C > 0$ in all directions.

Let us reiterate how that happened. If we only knew about the function $G_1(x_1, x_2)$, which is supported within the first x-quadrant, we could use the contours (5.3.28) to conclude that τ can be turned on to make $\mathcal{F}_{\Phi,\tau(x)}(x_1, x_2)$ decay in all directions *outside* the first quadrant. We would not deduce any decay within the first quadrant.

Similarly, if we only knew about $G_2(x_1, x_2)$, we could prove exponential decay of $\mathcal{F}_{\Phi,\tau(x)}(x_1, x_2)$ for suitable τ in any direction outside the third quadrant:



(5.3.32)

The magic of the argument is that since the two functions $G_1(p_1, p_2)$ and $G_2(p_1, p_2)$ agree on R, the two functions $\mathcal{F}_{\Phi,\tau}(x_1, x_2)$ defined by (5.3.23a) are in fact the *same*, which hence must decay in every direction!

The final step is to evaluate the inverse transform (5.3.23b) and determine for which complex (p_1, p_2) it converges. The main claim is that it does so for any sensible p which lies on a contour $\Gamma_{\Phi,\tilde{q}}$ where $\tilde{q}$ is in the convex hull of the admissible set (i.e., (5.3.31)) [21]. This region can be plotted and looks like a slightly distorted four-ball with $|p| \lesssim 1$.

In comparison with other approaches to the edge-of-the-wedge theorem, an advantage of this method (stressed in [21]) is that the original functions $G_i(p_1, p_2)$ need not be analytic in any open cone, and isolated directions suffice. (For example, if we have a representation $G_3(x_1, x_2)$ that vanishes for $x_1 > 0$, then $G_3(p_1, p_2)$ is not necessarily an analytic function of two variables, we only know that it is analytic in p_1.) This feature is essential for combining the representations which appear in (5.3.8). Also, the fact that we have to replace the set of admissible $\tilde{q}$ by its convex hull means that the FBI also generalizes the so-called "tube theorems". This is quite impressive for a method which amounts to Stokes' theorem combined with single-variable contour deformations!

An intriguing feature is that the deduced domain of analyticity depends on the initial choice of $\Phi(p)$. This seems to be relatively unexplored. Although Gaußians suffice to prove local statements, the regions which arise in physics tend to be large regions defined by inequalities of the type $-p_I^2 \leqslant m_I^2$. One wonders if judicious choices of $\Phi(p)$ could grant us access to deeper regions of the complex plane, for example simplifying proofs of crossing and its generalizations.

References

1. J. Maldacena, S.H. Shenker, D. Stanford, A bound on chaos. J. High Energy Phys. **08**, 106 (2016) [1503.01409]
2. S. Caron-Huot, M. Giroux, H.S. Hannesdottir, S. Mizera, What can be measured asymptotically? J. High Energy Phys. **01**, 139 (2024) [2308.02125]
3. S. Caron-Huot, M. Giroux, H.S. Hannesdottir, S. Mizera, Crossing beyond scattering amplitudes, 2310.12199
4. S. Weinberg, *The Quantum Theory of Fields. Vol. 1: Foundations* (Cambridge University Press, 2005). https://doi.org/10.1017/CBO9781139644167
5. R. Omnès, *Démonstration des relations de dispersion*, Les Houches Lect. Notes, vol. 10 (1960), p. 317
6. D.A. Kosower, B. Maybee, D. O'Connell, Amplitudes, observables, and classical scattering. J. High Energy Phys. **02**, 137 (2019) [1811.10950]
7. S.W. Hawking, Particle creation by black holes. Commun. Math. Phys. **43**, 199 (1975)
8. J. Bros, H. Epstein, V.J. Glaser, Some rigorous analyticity properties of the four-point function in momentum space. Nuovo Cim. **31**, 1265 (1964)
9. J. Bros, H. Epstein, V. Glaser, A proof of the crossing property for two-particle amplitudes in general quantum field theory. Commun. Math. Phys. **1**, 240 (1965)
10. J. Bros, V. Glaser, H. Epstein, Local analyticity properties of the n particle scattering amplitude. Helv. Phys. Acta **45**, 149 (1972)
11. J. Bros, Derivation of asymptotic crossing domains for multiparticle processes in axiomatic quantum field theory: A general approach and a complete proof for $2 \to 3$ particle processes. Phys. Rep. **134**, 325 (1986)
12. S. Mizera, Crossing symmetry in the planar limit. Phys. Rev. D **104**, 045003 (2021) [2104.12776]
13. H. Lehmann, K. Symanzik, W. Zimmermann, Zur Formulierung quantisierter Feldtheorien. Nuovo Cimento **1**, 205 (1955)
14. D. Ruelle, Connection between wightman functions and green functions in p-space. Il Nuovo Cimento (1955–1965) **19**, 356 (1961)
15. H. Araki, N. Burgoyne, Properties of the momentum space analytic function. Nuovo Cim. **18**, 342 (1960)
16. F.M. Haehl, M. Rangamani, *Records from the S-Matrix Marathon: Schwinger-Keldysh Formalism* (2024). 2410.10602
17. R. Britto, C. Duhr, H.S. Hannesdottir, S. Mizera, *Cutting-Edge Tools for Cutting Edges*, 2402.19415.
18. G. Sommer, Present state of rigorous analytic properties of scattering amplitudes. Fortsch. Phys. **18**, 577 (1970)
19. M. Correia, H.S. Hannesdottir, G. Isabella, A.M. Wolz, Z. Zhou, M. Giroux, S. Mizera, C. Pasiecznik, *Records from the S-Matrix Marathon: Gravitational Physics from Scattering Amplitudes* (2024), 2412.11649
20. F. Bloch, A. Nordsieck, Note on the radiation field of the electron. Phys. Rev. **52**, 54 (1937)
21. J. Bros, D. Iagolnitzer, Causality and local analyticity - mathematical study. Ann. Inst. H. Poincare Phys. Theor. **18**, 147 (1973)

A Timeless History of Time

6

Mang Hei Gordon Lee and Enrico Pajer

Abstract

By directly probing the initial conditions of our universe, cosmological surveys offer us a unique observational handle on quantum field theory in curved spacetime with dynamical gravity and might even allow us to glean information about a full theory of quantum gravity. Here we report on recent progress to study the natural observables in the problem, namely cosmological correlators. After setting the stage, we review results from three different approaches. First, we present the in-out formalism as an interesting alternative to the well-known in-in formalism and stress some of its advantages, such as the derivation of recursion relations, correlators cutting rules and a proposal for a de Sitter scattering matrix. Second, we tackle the important open problem of constructing effective theories in curved spacetime, which generally requires an open quantum system approach. Third, we provide an executive summary of general properties of the field-theoretic wavefunction that follow from symmetries, unitarity, causality and locality. We describe how these properties can be leveraged to bootstrap all tree-level results and we discuss loop contributions.

M. H. G. Lee
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK

Leung Center for Cosmology and Particle Astrophysics, National Taiwan University, Taipei, Taiwan
e-mail: mhglee@ntu.edu.tw

E. Pajer (✉)
Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

N. Arkani-Hamed et al. (eds.), *Records from the S-Matrix Marathon*,
Lecture Notes in Physics 1041, https://doi.org/10.1007/978-3-031-90352-6_6

6.1 Five Things All High-Energy Physicists Should Know About Cosmology

Enrico Pajer

In the following, we outline five facts about cosmology that all high-energy theorists should know about. These facts highlight the key role that early-universe cosmology plays in the quest to further our understanding of fundamental physics. We believe that these observations provide a strong motivation for high-energy theorists to care about cosmology.

6.1.1 *Fact 1*: All Cosmological Perturbations Are Primordial

One fact about our Universe is that perturbations in the distribution of everything that we observe on cosmological scale are primordial in origin. They can be causally generated only *before* the hot big bang.[1]

It is an experimental fact that the density of the Universe is slightly inhomogeneous at very early times. We see the consequences of this in the distribution of baryonic and dark matter and in the distribution of photons, the Cosmic Microwave Background (CMB). There are in fact several effects that link the inhomogeneities of the matter distribution to the CMB anisotropies. Figure 6.1 shows the correlation between temperature and polarization of the CMB, very roughly corresponding to density and velocity of the plasma respectively. On very large scales, at $l < 50$, we notice a clear detection of an anti-correlation. This is direct evidence that cosmological perturbations are coherent on those scales [1]. If one assumes the standard hot big bang, these scales are found to be out of causal contact with each other at the time the CMB was released. This shows that a *causal* explanation of initial conditions must be "primordial", i.e. pertaining to *before* the universe became filled with a hot thermal bath of particles.

6.1.2 *Fact 2*: Primordial Perturbation Are Adiabatic

On large scales primordial perturbations are all proportional to a *single* function of the spatial position $\mathbf{x}$:

$$\delta_a(\mathbf{x}, t) \equiv \frac{\delta\rho_a(\mathbf{x}, t)}{\bar{\rho}_a(t) + \bar{p}_a(t)} \propto \zeta(\mathbf{x}) \qquad \forall\, a\,, \tag{6.1.1}$$

[1] By hot big bang we mean the phase during which the Universe is filled with a hot bath of relativistic standard model particles.

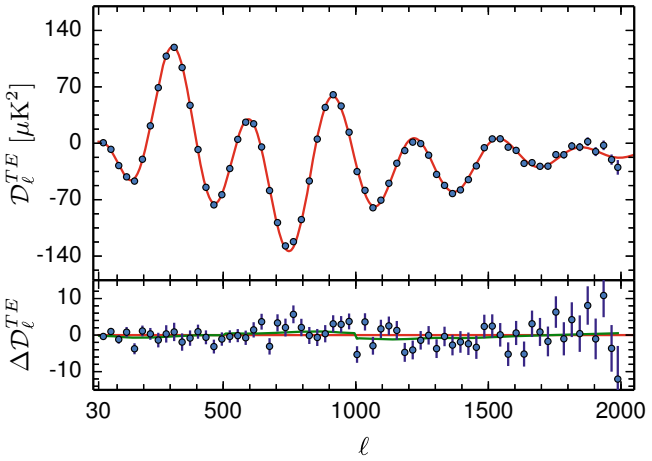


Fig. 6.1 The plot shows the cross-correlation ($\mathcal{D}_\ell^{TE}$) between temperature and polarization anisotropies of the CMB as a function of ℓ, the multiple moment, as measured by the Planck satellite [2]. The multiple l is related to the angular scale in the sky by $\theta \approx \frac{180^\circ}{\ell}$. High ℓ values correspond to smaller angular scales (fine details) and vice versa. The curve shows how the cross correlation varies with this scale. The red dots with blue error bars represent the actual measurements, whereas the solid line is the theoretical prediction based on the best-fit cosmological model. The anti-correlation around $l \lesssim 50$ is evidence that cosmological perturbations can be causally generated only before the hot big bang. The residuals (bottom plot) show that the differences between the observed data and the theoretical model are small at reasonably large distances. Reprinted under CC-BY-4.0 license from [2]. © 2016, The Author(s)

where ρ_a denotes the density and p_a the pressure (barred quantity denotes spatial average) of particles of kind a. Here, a accounts for everything[2] (baryon, dark matter, neutrinos, photons, etc.). For initial cosmological conditions, (6.1.1) is experimentally true to the percent (1%) accuracy level, as demonstrated by the tight bounds on isocurvature perturbations [3]. At face value, this suggests that the early Universe is not *that* messy.

This observation gives rise to one of the biggest open questions in cosmology: Why is this so simple? There are essentially two paradigms: (i) single-field inflation, which guarantees everything follows this a single "clock", and (ii) multifield inflation followed by a thermalization epoch that makes the many degrees of freedom decay into a single one.

[2] Incidentally, also the velocities δu_a of all these fluids are observed to be linearly proportional to ζ, as demonstrated by the tight bounds on "velocity" isocurvature perturbations. The precise expression on superHubble scales during radiation domination is $\delta_a(\mathbf{k}, t) = k^2 t^2 \zeta(\mathbf{k})/a^2 = -9\delta u(\mathbf{k}, t)/(2t)$.

In a sense it is not just a distribution of all forms of matter to be the same, but also the distribution of space-time itself that is determined by the same function (in the comoving gauge)

$$g_{ij} = a^2 \mathrm{e}^{2\zeta(\mathbf{x})}\delta_{ij} \quad \text{such that} \quad \delta g_{ii} = 6a^2\zeta(\mathbf{x})\,, \tag{6.1.2}$$

where $a(t)$ is the FLRW scale factor, whose dynamics is dictated by the Einstein's equations. Hence, to talk about cosmology we have to talk about quantizing the metric, at least perturbatively.

6.1.3 *Fact 3*: ζ Is Gaußian and Scale Invariant

Based on the previous paragraph, we better understand ζ as best as we can. The perturbations $\zeta(\mathbf{x})$ are statistically Gaußian to an accuracy better than 0.01% (this is one of the most precise results in cosmology [4]). This means that the relevant physics is characterized by its two-point function

$$\langle\zeta(\mathbf{x})\zeta(\mathbf{y})\rangle \sim \overset{\text{constant}}{\underset{\text{up to 3\%}}{}} \text{ or in Fourier space } \langle\zeta(\mathbf{k})\zeta(\mathbf{k}')\rangle = (2\pi)^3\delta(\mathbf{k}+\mathbf{k}')\frac{10^{-9}}{k^{3+(1-n_s)}}\,, \tag{6.1.3}$$

where n_s at the exponent is a the so-called *spectral tilt*, for which the current best measurement gives $n_s = 0.9649 \pm 0.0042$ [5]. The fact that the two-point function depends only very weakly on distance in position space, or equivalently that the power spectrum scales approximately as $\langle\zeta(\mathbf{k})\rangle \sim k^{-3}$ is known as the (approximate) *scale invariance* of primordial perturbations.

Scale invariance gives us a hint about what the spacetime looked like during the generation of primordial perturbations. The idea is that scale invariance emerged from a spacetime isometry. The deal is that if we want a homogeneous and isotropic spacetime and on top we demand for this extra "scale invariance" symmetry, we get a maximally symmetric spacetime called de Sitter spacetime with isometry group SO(4,1), corresponding to the conformal group in 3 Euclidean dimension. So in 1+3 space-time dimensions, the geometry of the primordial universe was well approximated by the de Sitter spacetime (in flat slicing):

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + \mathrm{e}^{2Ht}\mathrm{d}x^2 = \frac{-\mathrm{d}\eta^2 + \mathrm{d}x^2}{\eta^2 H^2}\,, \tag{6.1.4}$$

where $\eta = -\mathrm{e}^{-Ht}/H$, $-\infty < \eta < 0$, and where H is the unknown value of the Hubble parameter during inflation. In fact, it is "unknown" by some 37 orders of magnitude! The upper bound comes from the non-observation of primordial

gravitational waves and the lower bound from demanding that the universe reheats at a temperature above that of big bang nucleosynthesis,

$$10^{13}\,\text{GeV} > H \gg 10^{-24}\,\text{GeV}\,. \tag{6.1.5}$$

This is perhaps the second most unknown quantity in physics.

6.1.4 *Fact 4*: Large Scale Evolution Is Linear

Consider the distribution $\rho(\mathbf{x})$ of some substance and decompose it into a homogeneous background $\bar{\rho}(t)$ and some perturbations $\delta(t, \mathbf{x}) = (\rho - \bar{\rho})/\bar{\rho}$. Here δ could represent perturbations in the distribution of atoms, dark matter, photons, neutrinos, and so on. Then, on large scales we have

$$\delta(\mathbf{k}, t) = T^{(\delta)}(\mathbf{k}, t)\zeta(\mathbf{k})\,, \tag{6.1.6}$$

where $T^{(\delta)}$ denotes the so-called *transfer functions* for the perturbations δ. In general, $T^{(\delta)}$ depends on the substance we are looking at and on the choice of cosmology. The transfer functions are well known for the Λ-Cold Dark Matter (ΛCDM) model and can be computed numerically in a fraction of a second using a Boltzmann code such as CAMB.

6.1.5 *Fact 5*: Cosmological Observations Probe QFT in Curved Spacetime

We are finally coming to the punchline. On large scales, cosmological surveys measure QFT correlators of metric fluctuations

$$\underbrace{\left\langle \prod_{a=1}^{n} \delta(\mathbf{k}_a, t) \right\rangle}_{\text{from observations}} = \underbrace{\left[\prod_{a=1}^{n} T^{(\delta)}(\mathbf{k}_a, t) \right]}_{\text{known}} \underbrace{\left\langle \prod_{a=1}^{n} \zeta(\mathbf{k}_a) \right\rangle}_{\text{QFT in de Sitter}} + \underbrace{\text{non-linearities}}_{\mathcal{O}(\zeta^{n+1})}\,. \tag{6.1.7}$$

The central goal of primordial cosmology is to understand QFT and quantum gravity (if there is such a thing) asymptotically in de Sitter. So far we only have access to experimental data for $n = 2$, namely the primordial power spectrum (see e.g., Planck data [5]). The hope is that within the next 500 years we will have access to $n = 3, 4$ and, why not, maybe even higher. The prospects for detecting $\langle \zeta^3 \rangle$ are not so bad because it has a known lower limit [6], which is coming from the fact that gravity is a non-linear theory [7]. A recent discussion of future observational prospects can be found in [8].

If there is one formula to remember, (6.1.7) is the one.[3] It tells us why we should care about cosmology as field theorists and why we should consider anything beyond QFT. This represents the path forward.

6.2 The Ins and Outs of Cosmological Correlators

Enrico Pajer

The big picture is that we want to think about QFT in de Sitter because that allows us to compute the correlators in (6.1.7), and many different approaches to this central problem have been taken so far. In the last four or five years, many groups worldwide (recent work on the field theoretic wavefunction can be found, e.g., in [9–41]) have been exploring this problem from the perspective of the wavefunction of the Universe and its properties, and much of the progress in the field of the cosmological bootstrap has had the wavefunction itself as the main object of investigation. This is because the wavefunction is somehow a simpler and more primitive object: if we know the wavefunction of the Universe, we can compute whatever we want.

There is another approach, which is perhaps the oldest one: not computing the wavefunction but just computing the correlators. After all, these are the real observables. People have found that the computation of correlators is somewhat more involved for technical reasons, which we shall explain below.

6.2.1 The In-In Formalism

6.2.1.1 Setting Up the Stage

The system we are considering is that of a large semi-classical gravitational spacetime background that is very close to de Sitter space (and, in fact, for most of this section, we will work in exactly de Sitter space, although small perturbations can be described in perturbation theory). Later, we will allow quantum perturbations on top of it, so that the framework is that of quantum field theory on curved spacetime, which has a conformal diagram depicted as a square in Fig. 6.2.

The part of the square that is relevant for cosmology is the upper triangle and it is called the Poincaré patch. It is charted by the Poincaré coordinates

$$\mathrm{d}s^2 = \underbrace{-\mathrm{d}t^2 + a^2\mathrm{d}x^2}_{\text{FLRW coordinates}} = \underbrace{a^2(-\mathrm{d}\eta + \mathrm{d}x^2)}_{\text{conformal coordinates}} , \tag{6.2.1}$$

where $a = \mathrm{e}^{Ht} = -1/(\eta H)$ with constant Hubble parameter H.

[3] The non expert may simply neglect the non-linearities and assumed the transfer functions are known in the standard model, however it should be mentioned that a large research effort in the cosmology community is devoted to better understand and model non-linearities and to predict the transfer functions for more general models beyond ΛCDM.

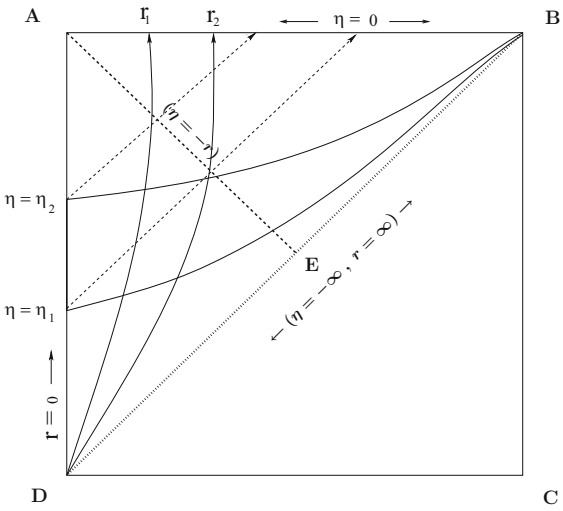


Fig. 6.2 Displayed is the complete Penrose diagram for de Sitter space-time, where each point represents a 2-sphere. The Poincaré patch labeled with ABD is described by the conformal chart (η, r, θ, ϕ). Line AB represents future infinity ($\mathcal{I}^+$), and lines BD (past null infinity where the Bunch–Davies vacuum is defined) and AE denote the cosmological horizon of a static observer at the south pole ($r = 0$). The diagram illustrates two constant-η spacelike hypersurfaces with $\eta_2 > \eta_1$, and two constant r, timelike hypersurfaces with $r_2 > r_1$. Dotted lines inclined at 45° show paths of gravitational waves emitted at $\eta = \eta_1, \eta_2$ on the worldline at $r = 0$, emanating from the source. The source is active during the time interval (η_1, η_2), emitting rapidly enough to fall within detectable frequencies. The AED region is static. Reprinted with permission from [42]. © 2016, American Physical Society. All rights reserved

In principle, if we knew everything about the Universe, we could find a full wavefunction that describes all degrees of freedom, including the metric. Provided we knew the initial conditions at the beginning of inflation in the infinite past (BD line in Fig. 6.2), and understood the theory of the Universe, we would be able to evolve it in time to find predictions on the AB line of the same figure. The AB line represents what, for de Sitter, is the future conformal[4] boundary and, in the context of inflation, represents the reheating surface where inflation transitions into the hot Big Bang. Let us emphasize once more that if we know the wavefunction on the AB line, then we essentially know everything about the Universe and can directly compute correlators on that future conformal boundary.

In this section, we will assume that the initial state is the Bunch–Davies vacuum. This state satisfies two important properties: (i) it is de Sitter invariant, and (ii) it asymptotes the Minkowski vacuum on very short scales. In principle, we could have started with a different initial condition, since at the level of QFT in curved

[4] The word "conformal" here is just a reminder that this "boundary" line can only be reached from a bulk point in an infinite proper time.

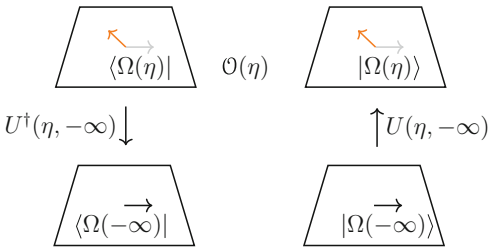


Fig. 6.3 On the bottom, we start with both a bra and a ket, which we evolve from minus infinity, where they are initially the Bunch–Davies states. After some time, we insert the operator that we want to use for a measurement. This is called an in-in observable because both the bra and the ket are prepared from the past. The arrows represent some vector in the Hilbert space and U denotes unitary evolution, which is geometrically just a rotation of the vectors

spacetime, other states in the Hilbert space are also allowed, albeit less motivated initial states.[5]

6.2.1.2 Correlators

While the wavefunction describes the entire system, what we actually measure, as mentioned earlier, are correlators. By correlators, we mean the expectation value of the product of local operators $\mathcal{O}$ on the state of the Universe Ω (presumably a vector in some Hilbert space of the QFT of everything), which, in momentum space, reads as

$$\lim_{\eta\to 0}\langle\Omega|\prod_{a=1}^{n}\mathcal{O}(\mathbf{k}_a,\eta)|\Omega\rangle \equiv \langle\Omega|\prod_{a=1}^{n}\mathcal{O}(\mathbf{k}_a)|\Omega\rangle \equiv \langle\mathcal{O}^n\rangle\,. \tag{6.2.2}$$

Here, η denotes a time variable. Most of the time, cosmologists are interested in equal-time products, and usually the time is taken to be the asymptotic future, $\eta \to 0$, because it is assumed that what we measure is the end result of inflation, since none of us was present during inflation to measure unequal-time correlators at finite η.

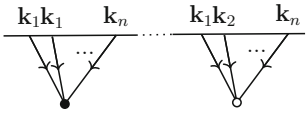
In practice, we know how to compute (6.2.2) in perturbation theory. The framework that makes this possible is known as the in-in formalism, in which we imagine evolving from the infinite past by turning on interactions adiabatically. More concretely, we start from the vacuum of the free theory and turn on the interaction slowly enough with an $i\varepsilon$ rotation of the time integral contour, and then evolve in time until some point (this generates the ket) where we insert an operator $\mathcal{O}_I(\eta)$, and then evolve back to generate the bra. This is summarized in Fig. 6.3 or in the following expression:

[5] An example would be the Bogliubov initial state, a set of states related to the Bunch–Davies vacuum by Bogliubov transformations [43,44]. See [35,45,46] for recent development.

$$\langle \mathcal{O}(\eta)\rangle = \langle 0|\Big[\overline{\mathcal{T}}\mathrm{e}^{i\int_{-\infty(1+i\varepsilon)}^{\eta}\mathrm{d}\eta' H_{\rm int}(\eta')}\Big]\mathcal{O}_I(\eta)\Big[\mathcal{T}\mathrm{e}^{-i\int_{-\infty(1-i\varepsilon)}^{\eta}\mathrm{d}\eta' H_{\rm int}(\eta')}\Big]|0\rangle\,, \tag{6.2.3}$$

in the interaction picture, which is computed order-by-order via Feynman diagrams (note that (6.2.3) is not an amplitude). This strategy was introduced about twenty years ago by Maldacena and then Weinberg, and has since been used by a variety of people to make calculations in a variety of models and to study formal properties of de Sitter. There are straightforward diagrammatic rules for calculations, which are reviewed in many references, for example [47, 48]. Here is an explicit example to illustrate what a typical calculation would require.

Example 6.2.1 We consider a contact diagram with one vertex $H_{\rm int} \sim \dot{\phi}^3$. To leading order in the coupling constant of this interaction, there are two diagrams to compute; one where the perturbations interact once on the bra and the other where the perturbation interacts on the ket. In this simple case, the two diagrams are just the complex conjugate of each other and so the sum can be written as a real part:



$$\begin{aligned}
&\overset{n=3}{=} 2\mathrm{Re}\Big[-i\lambda_1\int_{-\infty}^{0}\frac{\mathrm{d}\eta}{(H\eta)^4}(-H\eta)^3\prod_{a=1}^{3}G_r'(\eta,k_a)\Big]\\
&= 2\mathrm{Re}\Big[i\lambda_1\int_{-\infty}^{0}\frac{\mathrm{d}\eta}{H\eta}\prod_{a=1}^{3}\frac{H^2}{2c_sk_a^3}c_s^2k_a^2\eta\,\mathrm{e}^{ic_sk\eta}\Big]\\
&= 2\mathrm{Re}\Big[i\frac{\lambda_1H^5c_s^3}{8k_1k_2k_3}\int_{-\infty}^{0}\mathrm{d}\eta\,\eta^2\,\mathrm{e}^{iE_T\eta}\Big]\\
&= \frac{\lambda_1H^5c_s^3}{2k_1k_2k_3E_T^3}\,.
\end{aligned} \tag{6.2.4}$$

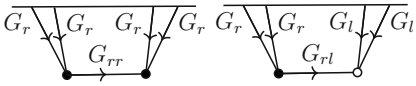
G_r' denotes the time derivative of the mode functions for a generic mass, which are Hankel functions. To go from the first to the second line, we assumed that $m = 0$ and used the massless mode functions for which

$$G(\eta,k) = \frac{H^2}{2c_sk^3}(1-ik\eta)e^{ik\eta}\,. \tag{6.2.5}$$

It is crucial to add both of these diagrams (according to (6.2.3), there are two Hamiltonians to keep track of, one inside an anti-time order and one inside a time ordering) to obtain a real correlator, which is necessary to ensure that the expectation value of a Hermitian operators is real. The three-point correlator computed above is typically what people search for in the sky in the form of non-Gaußianities in galaxy correlations, for example the so-called equilateral and orthogonal bispectrum templates (see [4] for recent bounds).

Although we know how to perform the above calculation given a Lagrangian, things get more complicated. (In this chapter, the notation of energies E and ω will be used interchangeably. Unless specified otherwise, both symbols mean the same thing.)

Example 6.2.2 For example, the next-to-the-simplest diagram involves a particle exchange. There are four diagrams, corresponding to picking two interactions $H_{\rm int}$ from left-left ($B^{(ll)}$), left-right ($B^{(lr)}$), right-left ($B^{(rl)}$) and right-right ($B^{(rr)}$) in (6.2.3). Since diagrams that differ from each other by exchanging all left and right vertices are complex conjugates of each other, we just need to compute two contributions



(6.2.6)

This example already gets algebraically involved in de Sitter so we discuss it in Minkowski, where the mode functions are simply

$$G(t,k)=\frac{e^{iEt}}{2E}\,, \tag{6.2.7}$$

with $E=\sqrt{k^2+m^2}$. The correlator is then found to be

$$\begin{aligned}
B^{(rr)}_{4,s} &= (-i\lambda)^2\int \mathrm{d}t_1\mathrm{d}t_2 G_r(t_1,k_1)G_r(t_1,k_2)G_{rr}(t_1,t_2,E_s)G_r(t_2,k_3)G_r(t_2,k_4)\,,\\
B^{(r\ell)}_{4,s} &= \lambda^2\int \mathrm{d}t_1\mathrm{d}t_2 G_r(t_1,k_1)G_r(t_1,k_2)G_{r\ell}(t_1,t_2,E_s)G_\ell(t_2,k_3)G_\ell(t_2,k_4)\,,\\
B_{4,s} &= B^{(rr)}_{4,s}+B^{(r\ell)}_{4,s}+B^{(\ell r)}_{4,s}+B^{(\ell\ell)}_{4,s}=2\mathrm{Re}\left[B^{(rr)}_{4,s}+B^{(r\ell)}_{4,s}\right]=\frac{2(E_T+E_s)\lambda^2}{E_L E_R E_T E_s\prod_a^4(2E_a)}\,,
\end{aligned} \tag{6.2.8}$$

where the partial energies are $E_R=E_3+E_4+E_s$ and $E_L=E_1+E_2+E_s$. We emphasize that *the number of integrals to perform for a single tree-level diagram increases exponentially in the number of vertices*: for a process involving V vertices, there are 2^V labeling of the diagram to consider, and each of them correspond to a different nested time integral over Hankel functions already at tree-level. Each diagram describes one way in which the interactions change the bra and/or the ket through time evolution. The increase in combinatorial complexity emerges from the need to keep track of both the right and left vertices in (6.2.3). Thus, the calculation quickly becomes intractable for a large enough V even though the integrated result is manifestly simple. In particular, the singularity structure of $B_{4,s}$ is physical while the individual $B^{(\bullet\star)}_{4,s}$ contributions have spurious poles that cancel in the sum. This indicates that doing the calculation in this way is probably not optimal.

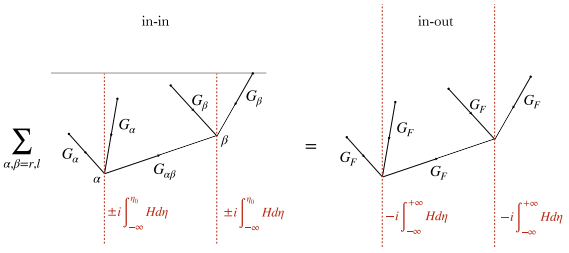
The formalism outlined above was not originally developed for cosmology, but was instead developed for entirely different reasons by Schwinger, and later by Keldysh, then by Feynman and Vernon, to calculate out-of-equilibrium quantum field theories or quantum field theories of open systems (e.g., Brownian motion of particles) rather than closed ones. This is known as the Schwinger–Keldysh formalism, where schematically for equal time correlators one encounters a path integral over a *closed time contour* where each field in the theory has been doubled,

$$\langle \phi^n \rangle = \int \mathrm{d}\phi\, \phi^n \int_{\rho_0}^{\phi} D\Phi_+ \int_{\rho_0}^{\phi} D\Phi_- \, \mathrm{e}^{i(S(\Phi_+)-S(\Phi_-)+F(\Phi_+,\Phi_-))} \,. \tag{6.2.9}$$

In cosmology, we use this formalism for closed systems (i.e., there are no fluctuations, no dissipation, no Feynman–Vernon influence $F(\Phi_+, \Phi_-) = 0$ and one assumes a pure state of a closed system)[6] in which we are trying to describe the whole Universe (all the degrees of freedom), so we never really allow this description to be applied to an open system. In the presence of an open quantum system, we further need to couple the two branches with $F(\Phi_+, \Phi_-) \neq 0$ of the path integral, and the situation becomes much more interesting. For the rest of this section we assume a close system and no interaction between the branches of the path integral (6.2.9), i.e. $F = 0$. The open system case will be discussed in Sect. 6.3.

Many different advances in the study of correlators have been made on various fronts, and we will describe some exciting progress below, starting with [48]. Before providing the full details, let us summarize the three main takeaway points:

- Cosmological correlators in de Sitter (and Minkowski) spaces can be computed using the in-in formalism (see Examples 6.2.1 and 6.2.2). We propose that there is a simpler method for such computations using the *in-out formalism*—namely, using only the familiar time-ordered (Feynman) propagators:



(6.2.10)

(Reprinted under CC-BY-4.0 license from [48]. © 2024, The Author(s).)
This method of computation is convenient because it eliminates the need to sum over vertices from the right and left time-evolution operators,

[6] This last assumption is an essential one for in-in = in-out.

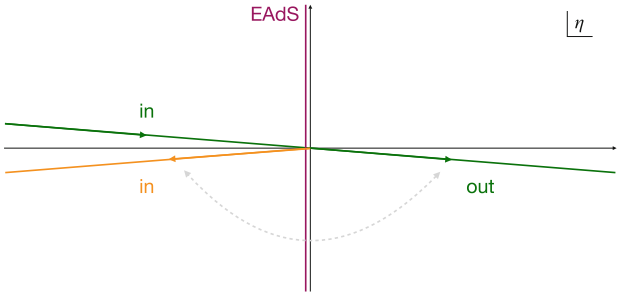


Fig. 6.4 The complex η (conformal time) plane with $-\infty < \eta < 0$ and the traditional path integral describe the theory as being performed over the yellow and green lines; these are the in-in contours of the path integral. These contours prepare the two indices of the density matrix of the system, allowing for the computation of any correlator. Reprinted under CC-BY-4.0 license from [48]. © 2024, The Author(s)

as it involves only one time-evolution operator. From the perspective of the integral, the only change is that the range of integration in η now extends from $-\infty$ to ∞. Because this approach leads to significant technical simplifications, we will use it in the context of cosmology to derive formal results that were previously far-fetched. One of our main goals, for example, will be to make the general consequences of unitarity, locality, and causality more manifest.

- For the equality in (6.2.10) to hold, some technical assumptions need to be made. One is that the interactions are IR finite. This is the case, for example, for single-field inflation, where interactions always have a sufficient number of derivatives[7] so that they become very soft in the future and thus go to zero as $\eta \to 0$, which is the future. We also assume that the evolution is unitary (a closed system in the Bunch–Davies vacuum). Moreover, we notice that the equality holds for any number of fields, regardless of their spin and mass.
- Finally, the in-out formalism leads to significant simplifications practically (many applications: new recursion relations, cutting rules, pole bagging) and conceptually (S-matrix technology, de Sitter S-matrix, non-perturbative optical theorem).

At this stage, let us note that there are different ways to interpret the above messages:

- The in-in time-evolution contour (green and yellow in Fig. 6.4), which lives exclusively at $\eta < 0$, can be deformed into an in-out contour (green)

[7] Non-derivative interactions such as ϕ^3 do appear for the inflaton, but not for the relevant quantity ζ.

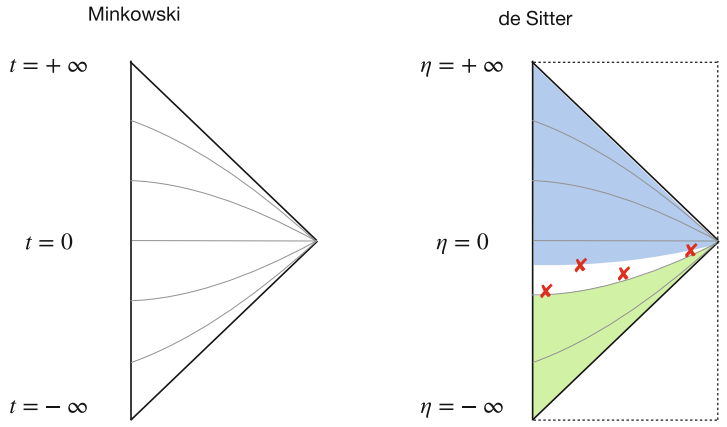


Fig. 6.5 Comparison between Minkowski (left) and de Sitter (right). On the right, the lower square is where the operators (red crosses) are inserted, and we aim to compute their expectation value. Typically, we prepare both the bra and the ket from the infinite past. Here, we highlight that the bra can be prepared from the infinite future of another copy of de Sitter (bra = blue region and ket = green region). As long as the bra is prepared to be the correct Bunch–Davies evolved, the in-in or in-out calculation yields the same result. The advantage of preparing the bra from the future is that it allows a single time ordering from $-\infty$ to ∞, resulting in a single (Feynman) propagator, which simplifies the technical process of obtaining the result. Reprinted under CC-BY-4.0 license from [48]. © 2024, The Author(s)

by adding a second spacetime (contracting Poincaré patch, where $\eta > 0$) that prepares the bra from the future and the ket from the past. Thus, in this picture, we have two different Poincaré patched of de Sitter spaces ($-\infty < \eta < 0$ and $0 < \eta < \infty$) that are merged together; one is an expanding de Sitter patch (on the left) and the other is a contracting de Sitter patch associated with positive conformal time instead of negative.

- The end result is a straight contour (green), just like in Euclidean anti de Sitter (vertical line) and in Minkowski amplitudes.

There is an conceptual advantage in the in-out formalism, as depicted in Fig. 6.5:

- Conceptually, the in-out formalism leads to a natural proposal for a de Sitter S-matrix, from a null past (null boundary of the green region) to a null future (null boundary of the blue region). This is also the case for massive particles.
- Unitarity-time evolution is imprinted in the de Sitter S-matrix via the standard optical theorem (both perturbative and non-perturbatively).

We now get to some more technical details.

6.2.2 In-In = In-Out

Let us slightly generalize the definition of an in-in correlator allowing for unequal time operators inside a time ordering[8]

$$B^{(n)}_{\text{in-in}} = \langle 0|\overline{\mathcal{T}}\left[\mathrm{e}^{i\int_{-\infty(1+i\varepsilon)}^{t_0}\mathrm{d}t\, H_{\text{int}}(t)}\right] \mathcal{T}\left[\mathcal{O}(t_1,\mathbf{x}_1)\dots\mathcal{O}(t_n,\mathbf{x}_n)\mathrm{e}^{-i\int_{-\infty(1-i\varepsilon)}^{t_0}\mathrm{d}t\, H_{\text{int}}(t)}\right]|0\rangle' , \tag{6.2.11}$$

where the prime removes the Dirac delta of momentum conservation. If we set $t_1 = \dots = t_n = t$, then we recover (6.2.3). We will refer to (6.2.11) as an in-in correlator. In this case, the $i\varepsilon$ rotation of the contour selects the Fock vacuum as the initial state (Bunch–Davies state) in the infinite past by adiabatically turning off interactions. Furthermore, t_0 is any time after all operator insertions.

Now, let us define an a priory different object, which we call an in-out correlator in de Sitter or Minkowski space:

$$B^{(n)}_{\text{in-out}} = \frac{\langle 0|\mathcal{T}\left[\mathcal{O}(t_1,\mathbf{x}_1)\cdots\mathcal{O}(t_n,\mathbf{x}_n)\mathrm{e}^{-i\int_{-\infty(1-i\varepsilon)}^{\infty(1-i\varepsilon)}\mathrm{d}t\, H_{\text{int}}(t)}\right]|0\rangle'}{\langle 0|\mathcal{T}\left[\mathrm{e}^{-i\int_{-\infty(1-i\varepsilon)}^{\infty(1-i\varepsilon)}\mathrm{d}t\, H_{\text{int}}(t)}\right]|0\rangle'} , \tag{6.2.12}$$

where we assume that the operators are inserted at negative conformal time (the standard expanding de Sitter patch). The denominator removes the vacuum-to-vacuum bubbles such that $\langle \mathbb{1}\rangle_{\text{in-out}} = 1$.

Now, there is a single time ordering that goes from $-\infty < \eta < \infty$, which includes the standard expanding Poincaré patch ($-\infty < \eta < 0$) and an extra contracting Poincaré patch ($0 < \eta < \infty$). Furthermore, we note that the $i\varepsilon$ rotation of the contour turns off the interactions adiabatically at the past and future null boundaries (notice the different signs of the $i\varepsilon$'s).

The claim now is that, for all IR-finite interactions, for which the time integral converges around $\eta = 0$, we have

$$(6.2.11) = B^{(n)}_{\text{in-in}} = B^{(n)}_{\text{in-out}} = (6.2.12) \tag{6.2.13}$$

This is a known fact in Minkowski (see e.g. [49]). Here we claim that it applies to de Sitter too (and probably to any accelerated FLRW but we have not checked yet). We next provide a formal argument and some explicit checks.

[8] When discussing general results we use η and t interchangeably. Then, when we consider explicit calculations and check we will use t in Minkowski and η in de Sitter.

6.2.2.1 A Formal Argument

A formal argument relies on the observation that infinite time evolution changes the ground (Bunch–Davies) state only by a phase

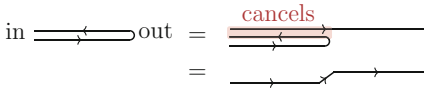
$$U(+\infty, -\infty)\,|0\rangle = |0\rangle\,\langle 0|U(+\infty, -\infty)|0\rangle\,. \tag{6.2.14}$$

Consequently, an infinite amount of time evolution from minus infinity to plus infinity does not result in a state excited with many particles, but rather returns us to the Bunch–Davies vacuum. This can be explicitly verified by projecting the left-hand side onto any n-point excited state and observing that the projection is zero for physical momenta.

Thus, our claim can be verified at all orders in perturbation theory by projecting onto any excited state. The result is a derivative of a delta function that enforces energy conservation, and which has zero support on physical perturbations. Then

$$\begin{aligned}
B_{\text{in-out}} &= \frac{\langle 0|T\left[\prod_{a=1}^{n}\phi(t_a)\,\mathrm{e}^{-i\int_{-\infty_-}^{+\infty_-}H_{\text{int}}\mathrm{d}t}\right]|0\rangle'}{\langle 0|T\left[\mathrm{e}^{-i\int_{-\infty_-}^{+\infty_-}H_{\text{int}}\mathrm{d}t}\right]|0\rangle'} \\
&= \langle 0|U^{\dagger}(+\infty_+, -\infty_+)T\left[\prod_{a=1}^{n}\phi(t_a)\,\mathrm{e}^{-i\int_{-\infty_-}^{+\infty_-}H_{\text{int}}\mathrm{d}t}\right]|0\rangle' \\
&= \langle 0|\overline{\mathcal{T}}\left[\mathrm{e}^{+i\int_{-\infty_+}^{+\infty_+}H_{\text{int}}\mathrm{d}t}\right]\mathcal{T}\left[\mathrm{e}^{-i\int_{t_0}^{+\infty_-}H_{\text{int}}\mathrm{d}t}\right] \\
&\quad \mathcal{T}\left[\prod_{a=1}^{n}\phi(t_a)\,\mathrm{e}^{-i\int_{-\infty_-}^{t_0}H_{\text{int}}\mathrm{d}t}\right]|0\rangle' \\
&= \langle 0|\overline{\mathcal{T}}\left[\mathrm{e}^{+i\int_{-\infty_+}^{t_0}H_{\text{int}}\mathrm{d}t}\right]U^{\dagger}(+\infty, t_0)U(+\infty, t_0) \\
&\quad \mathcal{T}\left[\prod_{a=1}^{n}\phi(t_a)\,\mathrm{e}^{-i\int_{-\infty_-}^{t_0}H_{\text{int}}\mathrm{d}t}\right]|0\rangle' \\
&= B_{\text{in-in}}\,,
\end{aligned} \tag{6.2.15}$$

where $\infty_{\pm} = \infty \pm i\varepsilon$. In terms of path integral contours, this is simply illustrated as follows:



(6.2.16)

Here, starting with the in-in contour, we consider adding an infinitely long contour for free, because acting on the vacuum amounts to just a phase. Then, the red-shaded parts of the contour cancel each other out, and we are left with the in-out contour, which gives the same result as the in-in one, up to a phase.

So, clearly, the formal argument holds, but now we would like to see how it works in practice.

6.2.2.2 Explicit Checks

There are many checks that we can perform. Here, we only go over a few of them (for more, see [48]) just to see how it works in perturbation theory.

We examine a basic theory involving a scalar field with a cubic interaction

$$H_{\rm int}(\eta) = \int_{\mathbf{x}} \frac{\lambda}{(n!)} F(\eta\partial_i, \eta\partial_\eta)\phi^n(\eta)\,, \tag{6.2.17}$$

where F represents a generic set of space and time derivatives acting on any of the fields, and ϕ can be any set of fields of any mass and spin. To simplify our discussion, we will focus on a single massive scalar. We will compute the n-point function to $\mathcal{O}(\lambda)$, with fields inserted at times $\eta_a \leqslant 0$ for $a = 1, \ldots, n$. In the context of the in-in formalism, and for simplicity setting $H = 1$, we find

$$\begin{aligned} B_{\text{in-in}} &= B^r_{\text{in-in}} + B^\ell_{\text{in-in}} \\ &= -i\lambda \int_{-\infty(1-i\varepsilon)}^{0} \frac{\mathrm{d}\eta}{\eta^4} F \prod_{a=1}^{n} G_F(\eta, \eta_a; k_a) \\ &\quad + i\lambda \int_{-\infty(1+i\varepsilon)}^{0} \frac{\mathrm{d}\eta}{\eta^4} F \prod_{a=1}^{n} G^+(\eta, \eta_a; k_a)^*\,. \end{aligned} \tag{6.2.18}$$

As long as the total number of η factors from F and the propagators exceeds four, the result converges at $\eta = 0$. This condition is met in scenarios like a local interaction with over three conformally coupled scalars, or interactions among massless scalars where $2n_{\partial_\eta} + n_{\partial_i} \geqslant 4$. In the in-out formalism, we find

$$B_{\text{in-out}} = -i\lambda \int_{-\infty(1-i\varepsilon)}^{+\infty(1-i\varepsilon)} \frac{\mathrm{d}\eta}{\eta^4} F \prod_{a=1}^{n} G_F(\eta, \eta_a; k_a)\,. \tag{6.2.19}$$

where G_F is the time-ordered propagator (the curved spacetime analog of the Feynman propagator). To verify that this is equivalent to the in-in expression, we

compute the difference. The portion of the in-out time integral from $-\infty$ to 0 precisely cancels out the right contribution of $B_{\text{in-in}}$, leaving us with

$$\begin{aligned} B_{\text{in-out}} - B_{\text{in-in}} &= -i\lambda \int_0^{+\infty_-} \frac{\mathrm{d}\eta}{\eta^4} F \left[\prod_{a=1}^{n} f_a(\eta) f_a^*(\eta_a) \right] \\ &\quad - i\lambda \int_{-\infty_+}^{0} \frac{\mathrm{d}\eta}{\eta^4} F \left[\prod_{a=1}^{n} f_a(\eta) f_a^*(\eta_a) \right] \\ &= -i\lambda \prod_{a=1}^{n} f_a^*(\eta_a) \int_{-\infty(1+i\varepsilon)}^{+\infty(1-i\varepsilon)} \frac{\mathrm{d}\eta}{\eta^4} F f_a(\eta) \,. \end{aligned} \tag{6.2.20}$$

Here, the label a on the mode function corresponds to the various momenta $\mathbf{k}_a$; it can also denote different fields, each with distinct mass values.

Because of the $i\epsilon$ deformations, the difference in (6.2.20) thus becomes an integral that can be closed in the lower half of the complex plane, where the integrand is analytic. According to Cauchy's theorem, this integral is zero, and we reach the desired conclusion. Higher-order diagrams are more complicated, but follow similar logic.

6.2.3 Applications

Next we discuss why this mathematical observation between in-in and in-out is *useful* in cosmology by considering a few concrete applications. These results are based on [48].

6.2.3.1 Pole Bagging

As previously mentioned, one reason why the correspondence between in-in and in-out is useful is because the in-out formalism involves the computation of fewer diagrams. Another example in which it is manifestly useful is the idea of "pole bagging." Indeed, a major simplification of the in-out formalism is that there is a single, time-ordered propagator, just like for amplitudes.

In Minkowski this is especially familiar in energy-momentum domain $(E, \mathbf{p}) = p^\mu$, where $G_F = 1/(p^2 + m^2 - i\varepsilon)$ (recall that we use the mostly plus signature) in which the time ordering is conveniently built in by the $i\varepsilon$ prescription. Thus, it is clear that $B_{\text{in-out}}$ are just the time-ordered correlators appearing in the LSZ reduction formulae for scattering amplitudes, i.e., $B_{\text{in-out}}$ are simply unamputated Feynman diagrams.

In cosmology, we tend to work in time-momentum domain $(t, \mathbf{k})$. The reason we do not Fourier transform time is that the background's expansion breaks time translation invariance and so different frequencies couple to each other already at linear order and energy/frequency is not conserved. Therefore, in Minkowski, this is simply a sum of poles, i.e., in-out correlators in Minkowski are computed by

repeatedly using

$$\int_{-\infty}^{\infty}\frac{\mathrm{d}\omega_i}{2\pi}\frac{1}{(\omega_i^2-\Omega_i^2)((\omega_i+\omega_X)^2-\Omega_j^2)}=-i\frac{\Omega_{ij}}{2\Omega_i\Omega_j(\omega_X^2-\Omega_{ij}^2)}\,, \tag{6.2.21}$$

where $\Omega_{ij}=\Omega_i+\Omega_j$ and $\Omega_a^2\equiv|\mathbf{k}_a|^2+M_a^2$.

Example 6.2.3 (Pole Bagging in Minkowski) For example, we consider the simplest case of a cubic polynomial interaction $\lambda\phi^3/3!$, where $F=\lambda$ and $n=3$. The Fourier transform of *equal-time* correlators then yields

$$\begin{aligned}B_3^{\rm flat}&=\int_{-\infty}^{\infty}\frac{\mathrm{d}\omega_1\mathrm{d}\omega_2\mathrm{d}\omega_3}{(2\pi)^3}(2\pi)\delta\left(\sum_{a=1}^{3}\omega_a\right)\mathrm{e}^{it\sum_{a=1}^{3}\omega_a}\lambda\prod_{b=1}^{3}\frac{1}{p_b^2+i\varepsilon}\\&=\int_{-\infty}^{\infty}\frac{\mathrm{d}\omega_1\mathrm{d}\omega_2}{(2\pi)^2}\frac{\lambda}{(\omega_1^2-\Omega_1^2+i\varepsilon)(\omega_2^2-\Omega_2^2+i\varepsilon)((\omega_1+\omega_2)^2-\Omega_3^2+i\varepsilon)}\,.\end{aligned} \tag{6.2.22}$$

Here, the first line is just the product of three Feynman propagators in the frequency domain, and we recall that the prime on the correlator (6.2.12) indicates that we have omitted $(2\pi)^3\delta^{(3)}\left(\sum\mathbf{k}_a\right)$.

The integrals can be evaluated using the residue theorem (see (6.2.21)), closing the contour in the upper half-plane. The first integral has poles at $\omega_1=-\Omega_1$ and $\omega_1=-\omega_2-\Omega_3$. This results in

$$B_3^{\rm flat}=-i\lambda\int\frac{\mathrm{d}\omega_2}{2\pi}\frac{\Omega_{13}}{2\Omega_1\Omega_3(\omega_2-\Omega_2)(\omega_2+\Omega_2)(\omega_2+\Omega_{13})(\omega_2-\Omega_{13})}\,, \tag{6.2.23}$$

where we have left the $i\varepsilon$'s implicit. It should be noted that there is a cancellation between two residues, which eliminates two of the zeros in the denominator (namely, $\omega_2=\pm(\Omega_1-\Omega_3)$), retaining only the zeros at $\omega_2=\pm\Omega_2$ and $\omega_2=\pm(\Omega_1+\Omega_3)$. The $i\varepsilon$ prescription directs us to consider only the two poles on the negative real axis, $-\Omega_2$ and $-\Omega_{13}$ (a strategy to maintain clarity in this calculation is to omit the $i\varepsilon$s in the integrand and introduce a slight counterclockwise rotation of the integration contours, ensuring that only the residues from poles on the negative real axis are picked up), which yields

$$B_3^{\rm flat}=-\frac{\lambda}{4\Omega_1\Omega_2\Omega_3(\Omega_1+\Omega_2+\Omega_3)}\,. \tag{6.2.24}$$

This yields the anticipated results from the bulk time integral, characterized by the simple E_T pole and the appropriate normalization factor for each $1/\Omega_a$. This can of course be done for any n-point function in Minkowski.

Below we illustrate how something very similar also holds in a de Sitter cosmology.

Example 6.2.4 (Pole Bagging in de Sitter) A similar approach works also in de Sitter for massless and conformally-coupled (c.c.) scalars using the above "dispersive" representation of the propagator. For example, we note that the Feynman (time-ordered) propagator of a massless or conformally coupled scalar in de Sitter can be written as

$$G_F^{\text{de Sitter}}(\eta,\eta',k)=\left(\frac{1}{2\pi i}\right)\int_{-\infty}^{\infty}\mathrm{d}\omega\frac{2\omega f_\omega(\eta)f_\omega^*(\eta')}{\omega^2-k^2+i\varepsilon}\,, \tag{6.2.25}$$

Hence, a similar computation as in Minkowski yields

$$\begin{aligned}
B_5^{\text{c.c.}} &= -iH^6\left(\frac{\eta_0}{2\pi i}\right)^5\int_{-\infty}^{\infty}\left(\prod_{i=1}^{5}\frac{\mathrm{d}\omega_i}{\omega_i^2-k^2+i\varepsilon}\right)\mathrm{e}^{i\omega_T\eta_0}\int_{-\infty}^{\infty}\mathrm{d}\eta\,\eta\,\mathrm{e}^{-i\omega_T\eta}\\
&= 2\pi H^6\left(\frac{\eta_0}{2\pi i}\right)^5\int_{-\infty}^{\infty}\left(\prod_{i=1}^{5}\frac{\mathrm{d}\omega_i}{\omega_i^2-k_i^2+i\varepsilon}\right)\mathrm{e}^{i\omega_T\eta_0}\delta'(\omega_T)\\
&= -2\pi i H^6\left(\frac{\eta_0}{2\pi i}\right)^5\\
&\quad\times\int_{-\infty}^{\infty}\frac{\mathrm{d}\omega_1\mathrm{d}\omega_2\mathrm{d}\omega_3\mathrm{d}\omega_4(\eta_0((\omega_T-\omega_5)^2-k_5^2))-2i(\omega_T-\omega_5))}{\left(\prod_{i=1}^{4}\left(\omega_i^2-k_i^2+i\varepsilon\right)\right)\left((\omega_T-\omega_5)^2-k_5^2+i\varepsilon\right)^2}\\
&= \frac{(H\eta_0)^6}{(k_1+k_2+k_3+k_4+k_5)(16k_1k_2k_3k_4k_5)}\,.
\end{aligned} \tag{6.2.26}$$

(Note that now, there is no delta function in the ω's because energy is not conserved in de Sitter.) One should be able to extend this to all masses in de Sitter, perhaps using embedding space coordinates, for which the propagators are simpler.

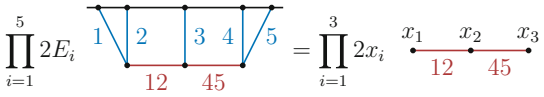
Let us now move on to a second application.

6.2.3.2 Recursion Relations

So far, the recursion relations we discuss below are found to work only in Minkowski. The key observation is that pole bagging has many similarities with integrations by part, and we can exploit this as follows. We first introduce *chains* to simplify correlators with many external lines, that is, we (schematically) write

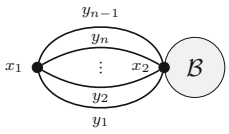
$$\text{correlator}=\frac{\prod_{i=1}^{m}2x_i}{\prod_{i=1}^{n}2E_i}\times\text{chains}\,. \tag{6.2.27}$$

Example 6.2.5 What we mean by (6.2.27) is, for example,



(6.2.28)

Here, the chains are computed by nested time integrals

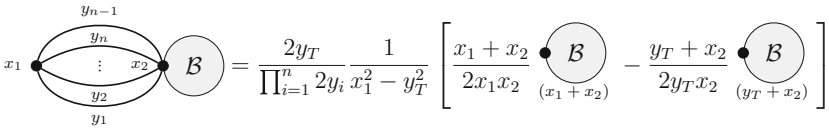


$$\equiv \int_{-\infty}^{\infty} \prod_{v\in\mathcal{B}\setminus\{2\}} i\,\mathrm{d}t_v\, G_F(t_v, t_0, x_t) \times \int_{-\infty}^{\infty} i\,\mathrm{d}t_2 G_F(t_2, t_0, x_2) \prod_{e\in\mathcal{B}} G_F(t_{v_e}, t_{v'_e}, y_{v_e}) I_n(\{y_i\}, t_2), \tag{6.2.29}$$

where $\mathcal{B}$ could be anything and with

$$I_n(\{y_i\}, t_2) \equiv \int_{-\infty}^{\infty} i\,\mathrm{d}t_1\, G_F(t_1, t_0, x_1) \prod_{i=1}^{n} G_F(t_1, t_2, y_i) = \frac{2y_T}{\prod_{i=1}^{n} 2y_i} I_1(y_T, t_2)\,. \tag{6.2.30}$$

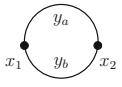
The t_1 integral can be done and re-written by explicitly performing one time integral we obtain the recursion relation as



(6.2.31)

The last equation tells us how diagrams with V vertices are related to those with $V-1$ vertices in a simple (rational) way. In practice, this tells us that what we need to compute at the end of the day are one-vertex diagrams. For a straightforward application, consider a one-loop exchange diagram.

Example 6.2.6 We have



$$= \frac{y_a+y_b}{2y_ay_b} \frac{1}{x_1^2-(y_a+y_b)^2} \left[\frac{x_1+x_2}{2x_1x_2} \frac{-1}{(x_1+x_2)^2} - \frac{y_a+y_b+x_2}{2(y_a+y_b)x_2} \frac{-1}{(y_a+y_b+x_2)^2} \right] = \frac{x_1+x_2+y_a+y_b}{4x_1x_2y_ay_b(x_1+x_2)(x_1+y_a+y_b)(x_2+y_a+y_b)}\,. \tag{6.2.32}$$

Let us mention that the recursions relations for correlators given above are very similar to the wavefunction recursion relations of [13]. For the wavefunction this relation becomes quite complicated at loop level, while it remains simple at the correlator level.

We now move on to the last application that we want to discuss.

6.2.3.3 Correlator Cutting Rules

We understand that unitarity for amplitudes is embodied in the optical theorem, which provides an explicit relationship between the imaginary part of the amplitude and the total cross-section in the forward limit. For correlators, a similar argument was not known until recently [48,50]. Here, we aim to demonstrate how representing the correlator as an in-out object enables us to apply the same derivation of the Cutkosky cutting rules to correlators, just as with scattering amplitudes. Along the way, we will emphasize all the small technical modifications required.

Our starting point is the field identity

$$\sum_{r=0}^{n}(-1)^r \sum_{\sigma\in\Pi(r,n-r)} \overline{\mathcal{T}}\left[\mathcal{O}_{\sigma(1)}...\mathcal{O}_{\sigma(r)}\right]\mathcal{T}\left[(\mathcal{O}_{\sigma(r+1)}...\mathcal{O}_{\sigma(n)}\right] = 0\,. \tag{6.2.33}$$

which may be recognized as Veltman's largest time equation [51] and it has already been discussed many times so far (see the contributions [52] and [53]). With $\Pi(r, n-r)$, we refer to the set of partitions of $\{1,\ldots,n\}$ into two subsets of sizes r and $n-r$. Thus, the sum involves 2^n terms. The fields $\mathcal{O}_i \equiv \mathcal{O}_i(t_i)$ represent arbitrary products of operators at the same time point. Our primary focus will be on scenarios where these operators are monomials in the fields of the theory and their derivatives.

As we shall see, (6.2.33) leads to infinitely many propagator identities, which in turn become correlator cutting rules. (These appear to be equivalent to the wavefunction cutting rules of [13], the advantage is that one works directly with observables, i.e. correlators, rather than with the more primitive wavefunction; perhaps a more explicit relation can be established along the lines of [36]) These identities are more easily derived when it is additionally assumed that all the operators in (6.2.33) are Hermitian, and the resulting equation is real, which may not be obvious from its original form. However, we can combine terms pairwise to rewrite it as

$$\begin{aligned}\sum_{r=0}^{n/2-1} \sum_{\sigma\in\Pi(r,n-r)} (-1)^r\, 2\,\mathrm{Re}\langle 0|\overline{\mathcal{T}}\left[\prod_{a=1}^{r}\mathcal{O}_{\sigma(a)}\right]T\left[\prod_{b=r+1}^{n}\mathcal{O}_{\sigma(b)}\right]|0\rangle \\ + \sum_{\sigma\in\Pi(n/2,n/2)} (-1)^{n/2}\,\mathrm{Re}\langle 0|\overline{\mathcal{T}}\left[\prod_{a=1}^{n/2}\mathcal{O}_{\sigma(a)}\right]T\left[\prod_{b=n/2+1}^{n}\mathcal{O}_{\sigma(b)}\right]|0\rangle = 0\,,\end{aligned} \tag{6.2.34}$$

for n even and as

$$\sum_{r=0}^{(n-1)/2} \sum_{\sigma\in\Pi(r,n-r)} (-1)^r \,\mathrm{Im}\langle 0|\overline{\mathcal{T}}\left[\prod_{a=1}^{r}\mathcal{O}_{\sigma(a)}\right] T\left[\prod_{b=r+1}^{n}\mathcal{O}_{\sigma(b)}\right]|0\rangle = 0\,, \tag{6.2.35}$$

for n odd.

Next, we use (6.2.33) to derive propagator identities. The initial step involves expanding the time-evolution operator within an in-out correlator up to a certain order in perturbation theory. We then consider the various powers of ϕ and $H_{\rm int}$ as the distinct operators featured in (6.2.33). This approach yields a set of identities. To illustrate this process, we first consider diagrams containing a single vertex and then those with two vertices.

Example 6.2.7 (Propagator Identities: Contact Diagrams) A diagram with one vertex contains a single power of $H_{\rm int}$ (i.e., there is only one interaction) and n instances of the field ϕ. As a simple example, consider the following choice

$$\mathcal{O}_1 = \phi(\mathbf{x}_1, t_0)^m\,, \quad \mathcal{O}_2 = \phi(\mathbf{x}_2, t_0)^{n-m}\,, \quad \mathcal{O}_3 = H_{\rm int}(t)\,. \tag{6.2.36}$$

Here, t_0 represents an arbitrary time. Since our analysis is confined to equal-time correlators, we will disregard the time dependence from now on. Assuming the operators in (6.2.36) are Hermitian, we can insert them into (6.2.35) to obtain the following position space expression

$$\mathrm{Im}\left\{\langle\mathcal{T}[\phi^n H_{\rm int}]\rangle - \langle\phi^m\mathcal{T}[\phi^{n-m}H_{\rm int}]\rangle - \langle\phi^{n-m}\mathcal{T}[\phi^m H_{\rm int}]\rangle\right\} \simeq 0\,. \tag{6.2.37}$$

Here, we have omitted the time-ordering or anti-time-ordering symbols when only a single time is involved. We used $\simeq 0$ to indicate that this identity is true only after performing time integration, due to the omission of the term $H_{\rm int}\mathcal{T}[\phi^n]$. This simplification is allowed because any term where a Hamiltonian interaction does not share the same time ordering as a field will integrate to zero. It would be nice to understand this in simple physical terms.

Now, each objects in (6.2.37) can be rewritten as a correlator and the result can be manipulated as follows after changing integration variables and using the properties of the Feynman propagator

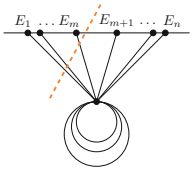
$$\begin{aligned} &B_n^{\rm c}(\{E_i\}_{i=1}^n) - \frac{1}{2}\left[B_n^{\rm c}(\{E_i\}_{i=1}^n) + (-1)^m B_n^{\rm c}(\{-E_i\}_{i=1}^m, \{E_i\}_{i=m+1}^n)\right] \\ &\quad - \frac{1}{2}\left[B_n^{\rm c}(\{E_i\}_{i=1}^n) + (-1)^{n-m} B_n^{\rm c}(\{E_i\}_{i=1}^m, \{-E_i\}_{i=m+1}^n)\right] = 0\,. \end{aligned} \tag{6.2.38}$$

Terms involving two sets have been analytically continued from respective positive to negative energies. Moreover, we observe that $B_n^{\rm c}(E_i{}_{i=1}^n)$ cancels out and if we subsequently invert the first m energies, we obtain

$$B_n^{\rm c}(\{E_i\}_{i=1}^n) + (-1)^n B_n^{\rm c}(\{-E_i\}_{i=1}^n) = 0\,. \tag{6.2.39}$$

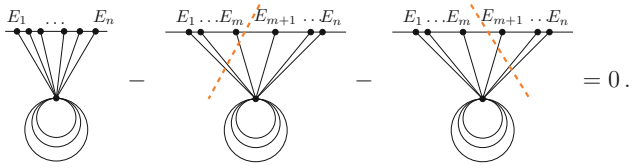
This is a necessary condition for this correlator to come from a unitary theory. It provides a useful constraint from unitarity, which operates directly at the level of observables. This is valuable from the standpoint of the cosmological bootstrap, which attempts to define good correlators based on their general properties rather than computing them from an explicit model.

There is a handy graphical notation for the cut identities correlators have to obey if the theory is unitary, e.g.,



$$= \frac{1}{2}\left[B_n^{\rm c}(\{E_i\}_{i=1}^n) + (-1)^m B_n^{\rm c}(\{-E_i\}_{i=1}^m, \{E_i\}_{i=m+1}^n))\right]. \tag{6.2.40}$$

The orange dashed line represents the energies that need to be analytically continued. The contact identity (6.2.37) thus becomes

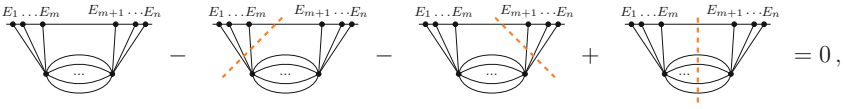


(6.2.41)

where the diagram without a cut should be interpreted as having a cut extending all the way to the right, or equivalently, to the left. It is important to note that this relation holds for the loop integrand at all loop orders, but only when considering a single vertex. In other words, we can contract any number of pairs of fields in $H_{\rm int}$ because they remain associated with the same time point in the largest time equation.

This works for more complicated diagrams.

Example 6.2.8 (Propagator Identities: Exchange Diagrams) For example, we can consider the exchange graph with melonic exchange



(6.2.42)

which reads, in formulae, as

$$B_n^{\text{ex},s}(\{E_i\}_{i=1}^n) + (-1)^n B_n^{\text{ex},s}(\{-E_i\}_{i=1}^n) = 2\int_{\mathbf{p}_1\cdots\mathbf{p}_{L+1}} \frac{B_{m,L+1}^{\text{c,cut}}(\{E_i\}_{i=1}^m, \{y_i\}_{i=1}^{L+1}) B_{n-m,L+1}^{\text{c,cut}}(\{E_i\}_{i=m+1}^n, \{y_i\}_{i=1}^{L+1})}{\prod_{i=1}^{L+1} P(y_i)}\,. \tag{6.2.43}$$

We emphasize that we have not yet found a suitable combinatorial structure to formulate the most general correlator cutting rule.

6.2.4 Scattering in de Sitter

Scattering in de Sitter space is very exciting because it provides a non-perturbative notion of unitarity in de Sitter space. We believe that this could be used to derive positivity bounds for cosmology within the next few years, indicating which effective field theories used in cosmology admit a consistent UV completion. The discussion below is based on [48], but see also [54–56] for other discussions of a de Sitter S-matrix. See also [57–64] for discussions of constraints on boost breaking effective field theories.

The in-out formalism we have developed offers us a clear strategy to build a scattering amplitude: take the in-out correlator and amputate its external legs with an LSZ-like projection. The result of this procedure automatically makes us think of an object similar to an S-matrix. Alternative, we can try to define a scattering matrix independently of correlators, while still using the in-out formalism. More formally, if we assume that we start with a state $|n\rangle$ in the infinite past ($\eta = -\infty$) that has n-particles

$$\underbrace{|n\rangle = \bigotimes_a^n |\Delta_a, \mathbf{k}_a, s_a, \sigma_a\rangle}_{\substack{\text{In and out states are tensor} \\ \text{products of unitary de Sitter irreps}}}\,, \tag{6.2.44}$$

(classified by the first Casimir/scaling dimension, momentum and spin) then we evolve for an infinite amount of time (from $\eta = -\infty$ to $\eta = +\infty$) and subsequently project onto a state with n' particles:

$$S_{n,n'} = \langle n' | U(+\infty, -\infty) |n\rangle = \langle n' | \mathcal{T} e^{-i\int_{-\infty}^{+\infty} H_{\text{int}}(\eta) d\eta} |n\rangle\,, \tag{6.2.45}$$

This is exactly what we would do in Minkowski, but in our formalism we can also do it in de Sitter.

Of course, people have discussed amplitudes in de Sitter space for around 20 years (see, e.g., [54]), and there is a variety of criticism regarding the existence of scattering amplitudes in de Sitter space. Let us address some of them from our setup:

- *IR divergences prevent a de Sitter S-matrix*. Possibly this is an issue, but we assume with a sufficient number of derivatives and/or in the presence of massive field the IR divergences do not arise.
- *Particles are unstable, so there are no asymptotic states*. Our $i\varepsilon$ prescription turns on/off interaction adiabatically at $\eta = \pm\infty$. Therefore, at least in perturbation theory we don't observe any dramatic instability.
- *Blue-shifted particles near null infinity (the "Big Bang") lead to a large backreaction*. This is a coordinate artifact. In global coordinates, a particle can cross the "Big Bang" null hypersurface that bounds the Poincaré patch from the past.
- *Particle creation prevents an out state at* $\eta = 0$. Possibly, but we work at $\eta = \pm\infty$.

Now that we have defined an S-matrix, let us compute it and see if the results are interesting. We define a (scalar) one-particle state in the relativistic normalization using

$$|\Delta, \mathbf{k}\rangle = \sqrt{2|\mathbf{k}|}a^\dagger_{\mathbf{k}}|0\rangle \ . \tag{6.2.46}$$

We define the connected part of the amplitude to be

$$\langle f|\, U(+\infty, -\infty) - 1\,|i\rangle = i(2\pi)^4\delta^{(3)}(\mathbf{k}_i - \mathbf{k}_f)A_{if} \ , \tag{6.2.47}$$

where we do not factor out the "energy conserving" delta function because energy is not conserved in de Sitter. Indeed as we will see $A_i\, j$ will turn out to be proportional to derivatives of a delta function. An interesting property that the answers will exhibit is a display of unitarity at the non-perturbative level, a fact which we check below in perturbative examples.

Example 6.2.9 Let's compute the simplest process: contact scattering of n conformally coupled scalars ($m^2 = 2H^2$) because they are very simple mode functions. Then to linear order in λ the result is

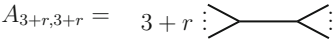
$$= A_{nn'} = -\lambda\,(-iH\partial_{E_T})^{n+n'-4}\delta(E_T)\,, \tag{6.2.48}$$

where E_T represents the total energy, accounting for the opposite signs of incoming and outgoing particles

$$E_T = -\sum_a^n |\mathbf{k}_a| + \sum_b^{n'} |\mathbf{k}_b| \,. \tag{6.2.49}$$

We found that the $(n+n'-4)$-th derivative of the energy Dirac delta ($n + n' = 4$ would therefore correspond to a Minkowski amplitude, which is consistent with the fact that ϕ^4 is a conformal interaction). It is interesting to note that, at least in perturbation theory, the S-matrix exhibits crossing symmetry. This allows us to consider all particles as outgoing (as also noted in [55]), so that E_T is the usual positive sum of norms.

As a more complex case, take the exchange diagram that enables the elastic scattering of $3 + r$ particles through the interaction $\lambda\phi^{4+r}/(4 + r)!$:

$A_{3+r,3+r} =$  . (6.2.50)

This diagram is infrared finite provided that $r \geqslant 0$, which explains the specific definition of r. A direct computation yields an expression much more complicated than (6.2.48):

$$A_{3+r,3+r} = \frac{\lambda^2 H^{2r}}{2} \sum_{l=0}^{r} b_l \frac{(k_{\rm in} - E_{\rm in})^{1+r-l} + (k_{\rm in} + E_{\rm in})^{1+r-l}}{2k_{\rm in}(-E_{\rm in}^2 + k_{\rm in}^2)^{1+r-l}} \partial^{r+l} \delta(E_{\rm in} - E_{\rm out}) \,, \tag{6.2.51}$$

where

$$E_{\rm in} = \sum_{a=1}^{3+r} |\mathbf{k}_a| \,, \qquad E_{\rm out} \equiv \sum_{a=4+r}^{6+2r} |\mathbf{k}_a| \,, \tag{6.2.52a}$$

$$\mathbf{k}_{\rm in} = \sum_{a=1}^{3+r} \mathbf{k}_a \qquad \mathbf{k}_{\rm out} \equiv \sum_{a=4+r}^{6+2r} \mathbf{k}_a \qquad b_l \equiv \frac{r!}{l!}(-1)^l \,. \tag{6.2.52b}$$

In particular, this involves a sum of terms each reminiscent of a Feynman propagator raised to some power with additional derivatives of a delta function.

The interesting claim now is that these de Sitter amplitudes obey the conventional optical theorem familiar from Minkowski QFT:

$$A_{i\to f} - A^*_{f\to i} = i \sum_X \int \mathrm{d}\Pi_X \, (2\pi)^4 \delta^{(3)}(\mathbf{k}_{\rm in} - \mathbf{k}_X) A_{iX} A^*_{fX} \,, \tag{6.2.53}$$

where the summation is over all potential states, with the sole distinction from Minkowski space being the retention of energy-conserving delta functions, thus eliminating the need to explicitly add any on the right-hand side. A significant implication of this theorem is that the right-hand side is clearly positive for forward scattering thanks to our symmetric definition of in and out states, which non-perturbatively restricts the imaginary part of A_{ii}! The hope is to one day use this as a way to obtain positivity bounds for de Sitter cosmology. The optical theorem is satisfied in a somewhat non-trivial manner, yet this can be verified through explicit examples.

Example 6.2.10 (An Explicit Check) We consider for example $4 \to 4$ scattering ($r = 1$). We have

$$\text{RHS of (6.2.53)} = i \int \frac{\mathrm{d}k_X^3}{(2\pi)^3} \frac{1}{2E_X} (2\pi)^4 \delta^{(3)}(\mathbf{k}_{\text{in}} - \mathbf{k}_X) |A_{4,1}|^2 \tag{6.2.54a}$$

$$= 2\pi i \lambda^2 H^2 \delta'(E_{\text{out}} - E_{\text{in}}) \frac{\delta'(E_{\text{in}} - \mathbf{k}_{\text{in}})}{2E_{\text{in}}} . \tag{6.2.54b}$$

and

$$\text{LHS of (6.2.53)} = i \frac{\lambda^2 H^2}{\mathbf{k}_{\text{in}}} \delta'(E_{\text{in}} - E_{\text{out}}) \,\text{Im}\, \frac{1}{(E_{\text{in}} - \mathbf{k}_{\text{in}} + i\text{e})^2} . \tag{6.2.55}$$

Admittedly, (6.2.54) and (6.2.55) appear quite different. However, integrating by parts and using the standard $i\varepsilon$ prescription to shift the energy pole results in a perfect match.

6.2.5 Outlook

In the near future, it would be interesting to:

- Use de Sitter amplitudes to derive positivity bounds. This requires understanding the analytic structure and asymptotics.
- Find an IR regulator that preserves in-in = in-out.
- Find a dispersive representation of the time-ordered propagator for any mass in de Sitter space to then reduce the correlator integrals to Minkowski amplitude integrals.
- Understand de Sitter amplitudes factorization.

6.3 Open Effective Field Theories

Enrico Pajer

The goal of primordial cosmology is to understand QFT and quantum gravity in (approximately, asymptotically) de Sitter spacetime. On large scales (≫Mpc) cosmological surveys measure QFT correlators of metric fluctuations as (see (6.1.7))

$$\left\langle \prod_{a=1}^{n} \delta\left(\mathbf{k}_a\right) \right\rangle \sim \left[\prod_{a=1}^{n} T^{(\delta)}\left(\mathbf{k}_a\right) \right] \left\langle \prod_{a=1}^{n} \zeta\left(\mathbf{k}_a\right) \right\rangle + \dots , \tag{6.3.1}$$

where the metric fluctuations $\delta\left(\mathbf{k}_a\right)$ are linked to observables such as CMB temperature fluctuations, density distribution of galaxies, dark matter, etc. and the curvature perturbations $\zeta\left(\mathbf{k}_a\right)$ are in de Sitter spacetime. In single-clock inflation models, the gravitational floor of non-Gaußianity in the primordial fluctuations is $f_{NL}^{eq} \gtrsim 10^{-2}$ [6], which provides a fundamental prediction to test against cosmological observations.

We have several aspirations regarding what fundamental physics we can learn about from cosmology. For example, we would like to:

- learn about new degrees of freedom and their interactions since inflation requires at least one degree of freedom and three energy scales beyond the standard model,[9]
- learn about the laws of gravity at short distances/high energies by probing GR and beyond at high energies,
- explore which QFTs are consistent in FLRW/de Sitter spacetime,
- and finally learn when QFT on curved space-time breaks down and quantum gravity is needed.

Here we develop a non-standard point of view on the physics of the primordial universe: we describe gravity plus the inflaton sector as an open system, remaining agnostic about other constients of the universe. We begin with an invitation to the relevant formalism, namely the Schwinger–Keldysh. Then we apply this description to general open effective field theories during inflation.

6.3.1 Invitation to the Schwinger–Keldysh Formalism

Here, we present an invitation to open quantum systems for high-energy physicists.

Instead of a pure state, in a general open system, we have to deal with a density matrix. This might seem unfamiliar or unnecessarily complicated; however, we find

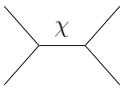
[9] These are: the energy scale of inflation, H, its first derivative, which is indirectly related to the duration of inflation and its second derivatives, which in the simplest models is measured in the tilt of the power spectrum.

it necessary to describe standard high-energy calculations, as we will show in this simply toy model. Consider the following Lagrangian for two scalars,[10] one of which is heavy χ and the other light ϕ

$$\mathcal{L}_{\rm UV} = -(\partial\phi)^2 - (\partial\chi)^2 - m^2\phi^2 - M^2\chi^2 - gM\phi^2\chi\,. \tag{6.3.2}$$

As a classic example of EFT, we have

$$\mathcal{A}_{2\to 2} \sim \quad \sim \frac{g^2M^2}{s-M^2} \sim g^2\sum_n \left(\frac{s}{M^2}\right)^n\,. \tag{6.3.3}$$



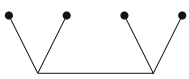
Then we can expand in local unitary, interactions

$$\mathcal{L}_{\rm eff} = (\partial\phi)^2 - m^2\phi^2 - g^2M^2\sum_n \phi^2\left(\frac{\Box}{M^2}\right)^n\phi^2 \tag{6.3.4}$$

at tree level, which depends only on M^2. Hence, when computing amplitudes, there is no need to use an open quantum system description.

What happens when we want to compute finite-time correlators? Let us consider the equivalent four-point correlator

$$B_4 = \quad = \langle\prod_{a=1}^4 \phi(t,\mathbf{k}_a)\rangle = \frac{(gM)^2}{E_S\prod_{a=1}^4 E_a}\frac{2(E_T+E_S)}{E_LE_RE_T}\,, \tag{6.3.5}$$



where we have

$$\begin{aligned} E_L &= E_1 + E_2 + E_S\,,\\ E_R &= E_3 + E_4 + E_S\,,\\ E_S &= \sqrt{|\mathbf{k}_1+\mathbf{k}_2|^2 + M^2}\,,\\ E_T &= \sum_{a=1}^4 E_a\,. \end{aligned} \tag{6.3.6}$$

Now we want to find the low-energy expansion of this as $M\to\infty$,

$$\lim_{M\to\infty} B_4^{\rm UV} = \frac{g^2}{\prod_{a=1}^4 E_a}\left[\frac{M^2-k_s^2+E_{12}E_{34}}{E_TM^2} + \frac{E_{12}E_{34}}{M^3} + \mathcal{O}(1/M^4)\right]\,, \tag{6.3.7}$$

[10] This Lagrangian appears often as a toy model of effective field theories. In this context, it was considered recently to discuss the field theoretic wavefunction [32].

where $E_{ij} = E_i + E_j$. This object is not Lorentz invariant, but this should not surprise us because we have chosen to look at a non-Lorentz invariant observable. More interestingly, the above result cannot come from a unitary EFT. Why? We just derived the cutting rules for correlators in the previous section. One of the rules was that $B_4(E) + (-1)^4 B_4(-E) = 0$, see (6.2.39), but the above formula is even in energy. This is related to the influence functional that we heard about in Ch. [52].

We can see by hand that the terms that are odd[11] in M do not come from a unitary EFT. The Lagrangian has to be a total derivative because otherwise it would contribute to the amplitude. A total derivative can also be written as a boundary term, and sees that the following would do the job,

$$\mathcal{S}_{\rm EFT} \supset \lambda \int d^3x \dot{\phi}^2 \phi^2 \Rightarrow B^{(4)}_{\rm EFT} = \quad = \frac{\sum_{a,b} E_a E_b E_T}{E_T \prod_{a=1}^4 E_a} \, {\rm Re}(i\lambda) \, . \tag{6.3.8}$$

However, to make this non-zero we need imaginary coupling $\lambda = iR$ which is non-unitary.[12] Thus, we can construct an EFT but not a unitary EFT. While this toy model does display some interesting features of full-fledged cosmological open systems, it has also some shortcoming and subtleties. In particular, the non-unitary terms are contact terms[13] and so arguably they could well be missed by the EFT. We nevertheless find it a useful example of how open system physics can show up in very familiar settings.[14]

The origin of the non-unitarity is that the quantum system is made of ϕ and χ but we are only observing the former. Since the two sectors are entangled in the interacting theory, tracing over one, say χ requires describing ϕ via the non-Hamiltonian evolution of a density matrix. Indeed, the correlator we are computing is schematically

$$\langle \phi^n \rangle = {\rm Tr}_{\phi,\chi}[\rho \phi^n] = {\rm Tr}_{\phi,\chi} \sum_{\chi'} \left[\rho |\chi'\rangle\langle\chi'| \phi^n \right] = {\rm Tr}_{\phi}[\rho_{\rm red} \phi^n] \, . \tag{6.3.9}$$

We had just a pure state, but after taking the trace over χ, the entanglement between ϕ and χ led to a mixed state. This situation in general requires to be studied with the Schwinger-Keldysh formalism, as we discuss below.

Audience Question 6.3.1 If we use the LSZ reduction formula, it will turn the correlator into the amplitude, for which we showed things worked well in a unitary local EFT. What's the deal?

[11] The original Lagrangian only depended on M^2 and so should the final result. In the above formulae, we have written $\sqrt{M^2}$ simply as M, assuming a real and positive mass.

[12] More in detail one can show the Hamiltonian is not Hermitian.

[13] E.P. thanks Alberto Nicolis for pointing this out.

[14] E.P. is in debt to Cliff Burgess, Alberto Nicolis, Thomas Colas and Santiago Agui Salcedo for useful discussions on this example.

Answer: Let us write the expression in Fourier space:

$$B_4^{\rm UV}(p_1, p_2, p_3, p_4) = \prod_{a=1}^{4} \frac{1}{p_a^2 + i\varepsilon} \frac{1}{(s - M^2)}\,, \tag{6.3.10}$$

In and EFT we would expand the off-sheel propagator as $\frac{1}{s-M^2} \simeq \frac{1}{M^2} \sum_n \left(\frac{s}{M^2}\right)^n$. This expansion however has a finite radius of convergence and so is valid only when $s < M^2$. When we take the Fourier transform to compute the correlator in the time-wavenumber domain, namely

$$B_4(t, \mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4) = \int {\rm d}p_1^0 {\rm d}p_2^0 {\rm d}p_3^0 {\rm d}p_4^0 B_4^{\rm UV}(p_1^0, p_2^0, p_3^0, p_4^0, \mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4)\,, \tag{6.3.11}$$

we necessarily integrate outside of the radius of convergence of the above Taylor expansion. Hence the result will not be correct if we expand before the Fourier transformation. Thus, working at fixed time forces us to integrate over energies outside of the radius of convergence of the EFT expansion, which here is $s = M^2$.

6.3.2 Open Quantum Cosmology

In general, more complicated systems arise from assembling simpler building blocks. In quantum mechanics, we build bigger Hilbert spaces $\mathcal{H}_{\rm tot}$ from tensor products of smaller Hilbert spaces, $\mathcal{H}_{\rm tot} = \otimes_i \mathcal{H}_i$. Open quantum systems arise when we observe only parts of $\mathcal{H}_{tot}$. For the Universe, this is always the case. When studying the cosmology of the very early universe there are many parts of the system that we do not observe: momenta conjugate of fields, spectator fields, gravitons, massive particles, etc. Hence, as a first step, we should always start with an open-system approach to cosmology, both at the quantum level and in the semi-classical limit.

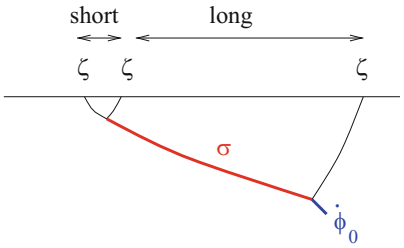
The defining characteristic of open systems is dissipation, which we use here as a catch-all word for a series of phenomena including fluctuations, non-unitary evolution, dispersion, etc. It is hard to give a general rule for when the presence of an environment is important and leads to sizable dissipative effects. A rule of thumb is the following: *"Open systems are needed when there are many "excited" degrees of freedom we do not observe or measure"*. We can gives some examples of cosmological scenarios that always or only sometimes require an open system approach:

- *Always:* particle production out of vacuum due to time-dependent background. In de Sitter, we expect that the "vacuum", as experienced by a

comoving observer, is an environment with a thermal bath at the Hubble temperature[15] $\mathrm{e}^{-E/T} \sim \mathrm{e}^{-2\pi m/H}$.

- *Sometimes:* physical production of particles, warm inflation, non-adiabatic kicks (features), etc. The size of the effect of the environment is model-dependent and can be large.
- *Sometimes:* long-short mode coupling, stochastic inflation. Again, the size of the effect is model dependent and can be large.

Let us look more closely at vacuum production in de Sitter that leads to curvature perturbations. The cosmological collider story considers the three-point function for the curvature perturbations ζ in the squeezed limit which in position space corresponds to two modes with a short separation distance k_{short}, and one mode at a long separation distance k_{long}, depicted as [65]:



(6.3.12)

(Reprinted with permission from [65]. © 2015, The Author(s). All rights reserved.) Here σ corresponds to a massive field that can decay into two inflatons, and is produced by the time-dependent background (de Sitter expansion) at a rate of $H \sim 2\pi T_d S$. This corresponds to corrections to the inflaton 3-pt function. The leading-order effects are captured by unitary evolution where there is no large dissipation. Hence, low-energy EFT is unitary but non-local, leading to characteristic non-analytic signals, as, for example, in the already mentioned squeeze limit of the three-point function B_3. This limit corresponds to small values of the ratio between the short and long separation distances. Considering, for example, the squeeze limit for a single particle with mass $m > \frac{3H}{2}$ and spin s, we have [65, 66]

$$\Delta_\pm = \frac{3}{2} \pm i\mu, \quad \mu = \sqrt{\frac{m^2}{H^2} - \frac{9}{4}}, \tag{6.3.13}$$

$$\frac{\langle \zeta\zeta\zeta \rangle}{\langle \zeta\zeta \rangle_{\text{short}} \langle \zeta\zeta \rangle_{\text{long}}} \sim \epsilon \mathrm{e}^{-\pi\mu} |c(\mu)| \left[\mathrm{e}^{i\delta(\mu)} \left(\frac{k_{\text{long}}}{k_{\text{short}}} \right)^{\frac{3}{2}+i\mu} + \mathrm{e}^{-i\delta(\mu)} \left(\frac{k_{\text{long}}}{k_{\text{short}}} \right)^{\frac{3}{2}-i\mu} \right] P_s(\cos\theta), \tag{6.3.14}$$

[15] The finite temperature comes about very explicitly when one restricts the description to the static patch of a given observer.

where Δ_i are conformal dimensions and ϵ is a slow roll parameter. This is a general picture for studying the spectrum of particles during inflation.

From here, we can consider warm inflation, wherein we can build a mechanism to continuously produce particles by draining energy from the background or inflaton. For example, a scalar (axion) coupled to $U(1)$ vector via $(\alpha/f)\phi F\tilde{F}$ leads to tachyonic production of A_i particles, which back react on perturbations

$$\left(\frac{\partial^2}{\partial\tau^2} + k^2 \mp 2aHk\xi\right) A_\pm(\tau, k) = 0, \qquad \text{with} \qquad \xi \equiv \frac{\alpha\dot{\phi}}{2fH}. \tag{6.3.15}$$

This induces, fluctuation, dissipation, and non-unitary evolution

$$\ddot{\phi} + 3H\dot{\phi} + \frac{dV}{d\phi} \simeq \frac{\alpha}{f}\langle \mathbf{E}\cdot\mathbf{B}\rangle. \tag{6.3.16}$$

The low energy EFT is non-unitary. However, since most of the gauge fields are produced around horizon crossing, the low-energy dynamics of the inflaton around the Hubble scale is necessary non-local if we integrate out the gauge fields. This theory does not have a small parameter to organize an EFT expansion in derivative and hence it must be treated by assuming a concrete UV model.

There are many other models of particle production. An interesting one considered in [67] (see Fig. 6.6) features the coupling between the inflaton ϕ and an additional degree of freedom χ, both minimally coupled to gravity,

$$\begin{aligned} S = \int \mathrm{d}^4x\sqrt{-g}\Big[&\frac{1}{2}M_{\rm Pl}^2 R - \frac{1}{2}(\partial\phi)^2 - V(\phi) - |\partial\chi|^2 + M^2|\chi|^2 \\ &-i\frac{\partial_\mu\phi}{f}\left(\chi\partial^\mu\chi^* - \chi^*\partial^\mu\chi\right) - \frac{1}{2}m^2\left(\chi^2 + \chi^{*2}\right)\Big]. \end{aligned} \tag{6.3.17}$$

In this case χ modes become tachyonic and are copiously produced. The big difference is that most modes are produced with a wavelength that is parametrically shorter than the Hubble radius. The χ environment leads to fluctuations, dissipation,

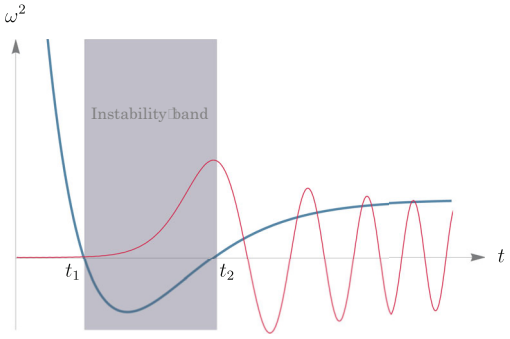


Fig. 6.6 The blue line depicts WKB frequency over time while the red line depicts the mode function with an amplitude that grows exponentially during the instability band. Reprinted with permission from [67]. © 2023, IOP Publishing Ltd and Sissa Medialab. All rights reserved. For further details, see the original source

etc. in the dynamics of ϕ, but crucially this can be described from the bottom up in terms of an open quantum system with a single degree of freedom and a *local* action! The corresponding linearized equation of motion in Fourier space takes the typical form of a Langevin equation,

$$\ddot{\varphi}_{\mathbf{k}} + (3H + \gamma)\dot{\varphi}_{\mathbf{k}} + \left(\frac{k^2}{a^2} + V''\right)\varphi_{\mathbf{k}} = -\frac{m^2}{f}\delta\mathcal{O}_S(\mathbf{k})\,, \tag{6.3.18}$$

where γ is the friction coefficient that causes dissipation and $\delta\mathcal{O}_S$ are fluctuations.

As a last example, we briefly discuss stochastic inflation [68, 69]. Sometimes the environment is not a different field, as in $\mathcal{H} = \mathcal{H}_\phi \otimes \mathcal{H}_\chi$ but different Fourier modes of the same field as in $\mathcal{H} = \otimes_{\mathbf{k}}\mathcal{H}_{\mathbf{k}}$. Then one ends up with a characteristic Langevin equation of the form

$$\frac{\mathrm{d}\bar{\varphi}}{\mathrm{d}t} = -\frac{V'}{3H} + \frac{H^{3/2}}{2\pi}\xi(t)\,. \tag{6.3.19}$$

where the second time derivative term has been dropped and the last term gives the fluctuations that short modes induce on long modes. This is typically treated non-perturbatively but semi-classically. In this case, there is a local, semi-classical, low-energy EFT.

We close this section by comparing bottom-up versus top-down approaches:

- Top-down pros: many explicit models exist which exhibit large open quantum system, a.k.a. "environmental", behavior.
- Top-down cons: calculations can be long and hard, it is difficult to find general patterns or phenomena, and models are delicate. By delicate we mean that small changes in the model can lead to intractable dynamics such as: non-locality in time, non-Markovian quantum dynamics, a breakdown of perturbation theory, instability, and large backreaction.
- Bottom-up pros: general features are visible, meaning that we can assume certain rules, such as stability, locality, perturbativity, etc. and we do not need to check the detailed implementation.
- Bottom-up cons: it is not clear whether one misses some constraints on the low-energy effective couplings and, relatedly, whether the model can be UV-completed.

Overall, our goal in what follows is to develop a systematic bottom-up approach to open quantum dynamics during inflation.

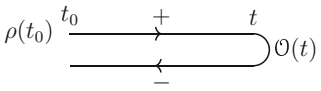
6.3.3 An Open Effective Field Theory of Inflation

The main idea is to use the principles of Effective Field Theory (EFT) to build large classes of bottom-up open quantum systems. The EFT rules are:

1. Name the low-energy *degrees of freedom*.
2. Choose *symmetries and principles* (e.g., locality) and write down the most generic symmetric action, which has infinitely many terms.
3. Choose a radiatively stable *power counting* and truncate the EFT to the desired precision with a *finite* number of operators.

We note that in the absence of locality in time and space, it is difficult to satisfy number 3 above.[16] The rest of this section follows closely [70] (see also [71] for a summary), which built upon previous results and especially [72,73].

We begin with number 1 above and consider the degrees of freedom of an open EFT. The simplest open EFT has a single degree of freedom, the Nambu–Goldstone boson of the spontaneous breaking of time translation by the inflaton background, as in the usual "unitarity" EFT of inflation. We use a close-time path integral, a.k.a. in-in formalism, which prepares a density matrix, as depicted here:



(6.3.20)

This is equivalent to the master equation known in operator language and has already been used for 20 years in inflation. However, hardly anyone has allowed for dissipation. It is convenient to perform the Keldysh rotation of the doubled fields, moving from the + and − contours of the path integral to the so-called retarded and advanced fields[17]

$$\begin{aligned} \pi_r &= \tfrac{\pi_+ + \pi_-}{2} \quad \text{and} \quad \pi_a = \pi_+ - \pi_- \\ \pi_+ &= \pi_r + \tfrac{\pi_a}{2} \quad \text{and} \quad \pi_+ = \pi_r - \tfrac{\pi_a}{2} \end{aligned} \tag{6.3.21}$$

We need to ensure our EFT satisfies several constraints, nicely summarise e.g., in [74,75]. We want our open QFT to come from a unitary "closed" UV theory with a Hermitian time evolution, so we have to satisfy

$$\begin{aligned} &S_{\text{eff}}\,[\pi_+,\pi_+] = 0\,, && S_{\text{eff}}\,[\pi_r,\pi_a=0] = 0\,, \\ &S_{\text{eff}}\,[\pi_+,\pi_-] = -S^*_{\text{eff}}\,[\pi_-,\pi_+]\,, && S_{\text{eff}}\,[\pi_r,\pi_a] = -S^*_{\text{eff}}\,[\pi_r,-\pi_a]\,, \\ &\text{Im}\,S_{\text{eff}}\,[\pi_+,\pi_-] \geqslant 0\,, && \text{Im}\,S_{\text{eff}}\,[\pi_r,\pi_a] \geqslant 0. \end{aligned} \tag{6.3.22}$$

[16] It is not impossible either: sometime one can Taylor expand in Fourier space around some fixed but non-zero momentum, as for the Fermi surface of fermions.

[17] These are also sometimes called "classical" and "quantum" fields. However the classical limit is subtle and we prefer to avoid this nomenclature.

These follow from conditions on a density matrix: normalisation requires $\mathrm{Tr}(\rho) = 1$, hermiticity means $\rho = \rho^\dagger$, and positivity means $\rho \geqslant 0$. Further details on the derivation of these constraints can be found in [70, 71].

Then we aim to write a general in-in functional, including a "unitary" part, $S_u(\pi_\pm)$, and a general "Feynman–Vernon" influence functional, $F(\pi_+, \pi_-)$, as

$$\begin{aligned} Z[J_\pm] &= \int D\pi_+ D\pi_- \mathrm{e}^{iS(\pi_+,\pi_-)+iJ_\pm\pi_\pm} \\ &= \int d\pi \int^{\pi} D\pi_+ \int^{\pi} D\pi_- \, \mathrm{e}^{i\left[S_u(\pi_+)-S_u(\pi_-)+F(\pi_+,\pi_-)+J_\pm\pi_\pm\right]} \\ Z[J_a, J_r] &= \int d\pi \int^{\pi} D\pi_r \int^{0} D\pi_a \, \mathrm{e}^{iS(\pi_r,\pi_a)+J_a\pi_a+J_r\pi_r} \,. \end{aligned} \tag{6.3.23}$$

In the Keldysh basis we see that π_a is non-dynamical because both its initial and final conditions are fixed by the boundaries of the path integral. It is only π_r that describes a dynamical field since its final condition is integrated over and hence arbitrary.

Now we address number 2 of our EFT rules. We need to ensure that our open QFT is consistent with general physical principles such as unitarity, locality, and causality. Unitarity of the UV theory is encoded in the three conditions already mentioned in the density matrix: $\mathrm{Tr}(\rho) = 1$, $\rho = \rho^\dagger$ and $\rho \geqslant 0$. We note that locality is not necessary because the environment can mediate interactions at a distance that are non-local in time (non-Markovianity). Nevertheless, we restrict ourselves to EFTs that are local in time and space, and hence we assume a separation of scales. These EFTs exist, as in the scalar warm inflation example. This is our strongest assumption, which leads to major simplifications. Finally, causality and analyticity are satisfied for the case we consider in the following, but additional constraints are expected to emerge from demanding causality of the UV completion as in amplitudes' positivity bounds [76].

Furthermore, we study the symmetries of our EFT. As in the EFT considered when addressing number 1, a general breaking of $\mathrm{U}_t(1)$ time translations is incompatible with scale invariance. We also need an internal shift symmetry $\mathrm{U}_{\mathrm{int}}(1)$ and we require a diagonal[18] is unbroken $\mathrm{U}_t(1) \times \mathrm{U}_{\mathrm{int}}(1) \to \mathrm{U}_{\mathrm{scale}}(1)$. This is depicted as

(6.3.24)

[18] Additionally one must assume that the degree of freedom enjoying the shift symmetry also evolves linearly, otherwise the low energy couplings depend on time, albeit in a fixed way [77].

The coset-construction for in-in is still relatively under developed (however, see [78, 79] for recent progress). Under time translations the Goldstone boson π_r transforms non-linearly, while the auxiliary advanced field π_a transforms linearly,

$$\begin{aligned}
&\pi_r(t,\boldsymbol{x}) \to \pi_r(t,\boldsymbol{x}) + \epsilon_r^0\,[1+\dot{\pi}_r(t,\boldsymbol{x})] + \mathcal{O}\left[\left(\epsilon_r^0\right)^2\right] \\
&\pi_a(t,\boldsymbol{x}) \to \pi_a(t,\boldsymbol{x}) + \epsilon_r^0\dot{\pi}_a(t,\boldsymbol{x}) + \mathcal{O}\left[\left(\epsilon_r^0\right)^2\right]
\end{aligned} \qquad \text{for} \quad \epsilon_+^0 = \epsilon_-^0 = \epsilon_r^0 \tag{6.3.25}$$

This tells us that π_a should not be interpreted as a Goldstone boson. Conversely, the other linear combination of time translations, which would transform fields as

$$\begin{aligned}
&\pi_r(t,\boldsymbol{x}) \to \pi_r'(t,\boldsymbol{x}) \\
&\quad = \pi_r(t,\boldsymbol{x}) + \frac{\epsilon_a^0}{2}\dot{\pi}_a(t,\boldsymbol{x}) + \mathcal{O}\left[\left(\epsilon_a^0\right)^2\right] \\
&\pi_a(t,\boldsymbol{x}) \to \pi_a'(t,\boldsymbol{x}) \\
&\quad = \pi_a(t,\boldsymbol{x}) + \epsilon_a^0\,[1+\dot{\pi}_r(t,\boldsymbol{x})] + \mathcal{O}\left[\left(\epsilon_a^0\right)^2\right]
\end{aligned} \qquad \text{for} \quad \epsilon_+^0 = -\epsilon_-^0 = \frac{\overline{\epsilon_a^0}}{2}\,. \tag{6.3.26}$$

is *explicitly* broken by the presence of the environment and plays no role in the following discussion. We will discuss non-linear boosts later in this section.

We now turn to the free theory. At quadratic order in perturbations and up to one derivative per field we have

$$\begin{aligned}
\mathcal{L}^{(2)} = {} & (\alpha_1 - 2\alpha_2)\,\dot{\pi}_r\dot{\pi}_a - \alpha_1\partial_i\pi_r\partial^i\pi_a - \alpha_0\pi_r\pi_a \\
& - 2\gamma_1\dot{\pi}_r\pi_a + i\left[\beta_1\pi_q^2 - (\beta_2-\beta_4)\,\dot{\pi}_q^2 + \beta_2\,(\partial_i\pi_a)^2\right].
\end{aligned} \tag{6.3.27}$$

The first line of $\mathcal{L}^{(2)}$ corresponds to unitarity dynamics[19] and the second line contains terms with dissipation γ_1 and fluctuations β_i. (Notice the imaginary unit i in front of the square brackets.) Two point functions are given by the Keldysh propagator G_K and the retarded/advanced propagators $G_{R,A}$ as

$$\begin{pmatrix} \langle\pi_r(x)\pi_r(y)\rangle & \langle\pi_r(x)\pi_a(y)\rangle \\ \langle\pi_a(x)\pi_r(y)\rangle & \langle\pi_a(x)\pi_a(y)\rangle \end{pmatrix} = \begin{pmatrix} iG_K(x,y) & -G_R(x,y) \\ -G_A(x,y) & 0 \end{pmatrix}. \tag{6.3.28}$$

[19] Unitary dynamics comes from $S(\pi_+) - S(\pi_-)$ and hence only generates terms that are odd in π_a.

Note π_a does not propagate, as anticipated. Here $G_A = G_R^*$ are the advanced and retarded Green's functions of the dissipative equations of motion, while G_K describes the perturbations in the system induced by fluctuations in the environment.

Next, we look more closely at propagators for this free theory. We can look to flat-space examples as illuminating warm ups:

$$G_{R/A}(k;\omega) = -\frac{1}{\omega^2 \pm i\gamma\omega - c_s^2 k^2} = -\frac{1}{(\omega_- - \omega_+)}\left[\frac{1}{\omega-\omega_-} - \frac{1}{\omega-\omega_+}\right],$$
$$\omega_\pm = -i\frac{\gamma}{2} \pm E_k^\gamma \quad \text{and} \quad E_k^\gamma = \sqrt{c_s^2 k^2 - \frac{\gamma^2}{4}}\,. \tag{6.3.29}$$

Here retarded/advanced propagators feel dissipation, which shifts poles in the complex plane but are independent of fluctuations, since $G_{A,R}$ are state independent. Dissipation can be seen in the presence of a complex pole, leading to exponential suppression in the power spectrum. This represents erasure of memory from the distant past due to dissipation into the environment.

The Keldysh propagator instead probes fluctuations

$$G_K(k;\omega) = -G_R(k;\omega)\widehat{D}_K(k;\omega)G_A(k;\omega) = -i\frac{\beta_1 + \beta_2\omega^2 + \beta_3 k^2}{\left(\omega^2 - c_s^2 k^2\right)^2 + \gamma^2\omega^2} \tag{6.3.30}$$

for which the final spectrum is largely specified by fluctuations

$$P_k = \frac{2\beta_1}{\gamma c_s^2 k^2} + \frac{2\beta_2}{\gamma} + \frac{2\beta_3}{c_s^2\gamma}\,. \tag{6.3.31}$$

Next, we consider de Sitter propagators. The mode functions are still Hankel functions, but now the index depends on dissipation. Indeed in de Sitter the Hubble expansion leads to a characteristic dissipative term proportional to the number of spatial dimensions. In the presence of dissipation due to an environment, γ can be thought of as a change in the number of dimensions. Indeed, the asymptotic time dependence of fields is found to be

$$\lim_{\eta\to 0}\pi_k(\eta) = \widetilde{A}\eta^{\frac{3}{2}} + \frac{\gamma}{2H}H^{(1)}_{\frac{3}{2}+\frac{\gamma}{2H}}(\eta)\,, \tag{6.3.32}$$

where $\pi \sim \eta^{\Delta}$ and $\Delta = \left(0, 3 + \frac{\gamma}{H}\right)$. Here, massless scalars still freeze out ($\gamma \geqslant 0$). The propagators are pretty messy, especially G_K (G_γ and F_γ are combinations of ${}_2F_3$, see [70]). We find

$$G^R(k;\eta_1,\eta_2) = -i\frac{\pi}{4}H^2(\eta_1\eta_2)^{\frac{3}{2}}\left(\frac{\eta_1}{\eta_2}\right)^{\frac{\gamma}{2H}} \operatorname{Im}\left[H^{(1)}_{\frac{3}{2}+\frac{\gamma}{2H}}(-k\eta_1)\,H^{(2)}_{\frac{3}{2}+\frac{\gamma}{2H}}(-k\eta_2)\right]\theta(\eta_1-\eta_2)\,, \tag{6.3.33}$$

$$\begin{aligned} G_1^K(k;\eta_1,\eta_2) = -i\frac{\widetilde{\beta}_1\pi^2}{8}(\eta_1\eta_2)^{\frac{3}{2}+\frac{\gamma}{2H}} \operatorname{Re}\Big\{ & H^{(1)}_{\frac{3}{2}+\frac{\gamma}{2H}}(-k\eta_1)\,H^{(1)}_{\frac{3}{2}+\frac{\gamma}{2H}}(-k\eta_2)\left[F_\gamma(z_2)-F_\gamma(\infty)\right] \\ & -H^{(1)}_{\frac{3}{2}+\frac{\gamma}{2H}}(-k\eta_1)\,H^{(2)}_{\frac{3}{2}+\frac{\gamma}{2H}}(-k\eta_2)\,G_\gamma(z_2)\Big\} + (\eta_1 \leftrightarrow \eta_2)\,. \end{aligned} \tag{6.3.34}$$

Symmetries ensure $P \sim 1/k^3$, so smoking-gun signals come from the bispectrum.

Now, we consider interactions for our open EFT. To cubic order in fluctuations, the in-in action is organized in powers of π_q.

$$\begin{aligned} \mathcal{L}_1^{(3)} =& (4\alpha_3 - 3\alpha_2)\,\dot{\pi}_r^2\dot{\pi}_a + \alpha_2(\partial_i\pi_r)^2\dot{\pi}_a + 2\alpha_2\dot{\pi}_r\partial_i\pi_r\partial^i\pi_a \\ & + (4\gamma_2 - \gamma_1)\,\dot{\pi}_r^2\pi_a + \gamma_1(\partial_i\pi_r)^2\pi_a\,, \\ \mathcal{L}_2^{(3)} =& i\left[-\beta_3\dot{\pi}_r\dot{\pi}_a\pi_q + \beta_3\partial_i\pi_r\partial^i\pi_a\pi_a + 2\beta_4\dot{\pi}_r\dot{\pi}_q^2 - 2\beta_4\partial_i\pi_r\partial^i\pi_a\dot{\pi}_a\right], \\ \mathcal{L}_3^{(3)} =& \delta_1\pi_a^3 + (\delta_5 - \delta_2)\,\dot{\pi}_a^2\pi_a + \delta_2(\partial_i\pi_a)^2\pi_a - \delta_4(\partial_i\pi_a)^2\dot{\pi}_a + (\delta_4 - \delta_6)\,\dot{\pi}_a^3\,. \end{aligned} \tag{6.3.35}$$

The real power of the EFT comes from relating operators at different orders because of the non-linearly realised boost, as noted early in the dissipative context in [72]. The inflaton background breaks both t-translation and boosts, where the former are resurrected by a diagonal combination with shift symmetry, and the latter are non-linearly realized ("broken").

We consider non-linear symmetries in the EFT. In the flat-space decoupling, the most general symmetry transformation is

$$\begin{aligned} \pi_r(t,\boldsymbol{x}) \to \pi_r'(t,\boldsymbol{x}) &= \pi_r\left(\Lambda^0_{r\mu}x^\mu, \Lambda^i_{r\mu}x^\mu\right) + \Lambda^0_{r\mu}x^\mu - t\,, \\ \pi_a(t,\boldsymbol{x}) \to \pi_a'(t,\boldsymbol{x}) &= \pi_a\left(\Lambda^0_{r}x^\mu, \Lambda^i_{r\mu}x^\mu\right)\,. \end{aligned} \tag{6.3.36}$$

Clearly π_r transforms non-linearly, hence relating $\mathcal{L}_2$ to $\mathcal{L}_3$, as already pointed out in [72, 73]. Once again notice that π_a transforms as a spectator field and is hence not a Goldstone boson. To proceed from here, we can write down all operators and demand the above symmetry, or work with invariant combinations. Using invariant combination we see the relation between quadratic and cubic orders in the Lagrangian

$$\begin{aligned}\mathcal{L}_1 = \mathcal{L}_1^{\rm LO} + \sum_{n=2}^{\infty} \gamma_n \left[-2\dot{\pi}_r + \left(\partial_\mu \pi_r\right)^2\right]^n \pi_a \\ - \sum_{n=2}^{\infty} \alpha_n \left[-2\dot{\pi}_r + \left(\partial_\mu \pi_r\right)^2\right]^{n-1} \left(-\dot{\pi}_a + \partial^\mu \pi_r \partial_\mu \pi_a\right) ,\end{aligned} \tag{6.3.37}$$

where we neglected higher-order terms with two or more derivatives per field.

We now turn to correlators. Correlators can be computed in perturbation theory using the standard in-in rules, being careful about distinguishing between retarded and Keldysh propagators. To build intuition, consider 3-point function in flat space which are mostly of the form [70] (see Fig. 6.7)

$$B_3 \sim \frac{\text{Poly}\,(E_1, E_2, E_3)}{\text{Sing}_\gamma} , \tag{6.3.38}$$

$$\begin{aligned}\text{Sing}_\gamma = \left|E_1^\gamma + E_2^\gamma + E_3^\gamma + \frac{3}{2} i\gamma\right|^2 \left|-E_1^\gamma + E_2^\gamma + E_3^\gamma + \frac{3}{2} i\gamma\right|^2 \\ \times \left|E_1^\gamma - E_2^\gamma + E_3^\gamma + \frac{3}{2} i\gamma\right|^2 \left|E_1^\gamma + E_2^\gamma - E_3^\gamma + \frac{3}{2} i\gamma\right|^2 ,\end{aligned} \tag{6.3.39}$$

$$E_k^\gamma = \sqrt{c_s^2 k^2 - \frac{\gamma^2}{4}} . \tag{6.3.40}$$

The 3-point function in flat space peaks when $k_i \pm k_j \pm k_l = 0$, namely for folded triangles The folded divergence is regulated by dissipation $\gamma \neq 0$, because interactions can build only over a finite amount of time before being erased by dissipation.

Cosmologists have used the in-in formalism for 20 years and yet the vast majority of papers assumed unitarity and non-dissipative dynamics. Here we have developed a systematic single-field open EFT for inflation, based on previous work by [72, 73]. Assuming locality in time and space, the resulting EFT is easy to write down and can be studied in perturbation theory. The smoking gun signal are peaks in the bispectrum near folded triangles. Our formalism is a starting point for studying aspects of quantum information of inflation such as entanglement growth, purity, decoherence, and more.

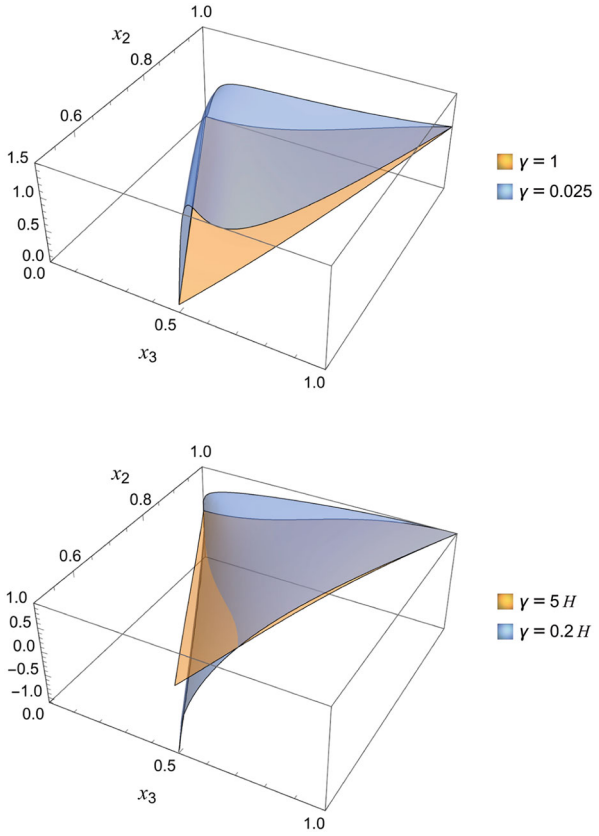


Fig. 6.7 Shape function $S(x_2, x_3) \equiv (x_2 x_3)^2[B(k_1, x_2 k_1, x_3 k_1)/B(k_1, k_1, k_1)]$ for the bispectrum corresponding to π_a^3 in Minkowski (*top*) and in de Sitter (*bottom*). Reprinted under CC-BY-4.0 license from [70]. © 2024, The Author(s)

6.4 Bootstrapping the Analytic Wavefunction

Mang Hei Gordon Lee

The purpose of this section is to provide a quick overview on how to bootstrap the cosmological wavefunction. We will introduce the off-shell wavefunction coefficients and explain how they can be constrained by physical principles such as unitarity, locality, and scale invariance. We will also illustrate this by bootstrapping the three-point wavefunction of a single massless scalar. This discussion is based on [18–21, 32, 80].

6.4.1 Cosmological Wavefunctions

The story begins with the Bunch–Davies vacuum state, $|\mathrm{BD}\rangle$, in the far past projected onto a field basis at time η_0 (usually $\eta_0 \to 0$, i.e., at the end of inflation)

$$\Psi[\phi_0, \eta_0] = \langle \phi_0, \eta_0 | \mathrm{BD} \rangle = \int_{\mathrm{BD}}^{\phi(\eta_0)=\phi_0} \mathrm{D}\phi \, \mathrm{e}^{i S[\phi;\eta_0]} \,, \tag{6.4.1}$$

where S denotes the action of the theory. From this we can determine the correlator of n (scalar) fields:

$$\langle \phi(\mathbf{k}_1) \dots \phi(\mathbf{k}_n) \rangle = \int \mathrm{D}\phi \, |\Psi(\phi)|^2 \, \phi(\mathbf{k}_1) \dots \phi(\mathbf{k}_n) \,. \tag{6.4.2}$$

To make the definition of the cosmological wavefunction in (6.4.1) more tractable, we usually expand it as follows [18, 81]

$$\Psi[\phi; \eta_0] = \exp\left[-\sum_{n=2}^{\infty} \int \prod_{i=1}^{n} \left(\frac{d^D \mathbf{k}_i}{(2\pi)^D} \right) \, \psi_n(\mathbf{k}_1, ..., \mathbf{k}_n; \eta_0) \prod_{i=1}^{n} \phi(\mathbf{k}_i) \right] , \tag{6.4.3}$$

where n denotes the number of vertices with external legs attached. The *on-shell* wavefunction coefficients ψ_j's, which are functions of external energies ($\omega_k = \sqrt{\mathbf{k}^2 + m^2}$) and spatial momenta ($\mathbf{k}$), are what we wish to bootstrap. What is meant by "bootstrap" will be clarified soon. Moreover, ω_k is the variable we will eventually continue in to go off-shell.

Let us mention that at tree level, correlators contain the same information as the wavefunctions, e.g., [18, 81]

$$\langle \phi(\mathbf{k}_1)\phi(\mathbf{k}_2)\phi(\mathbf{k}_3) \rangle \sim \frac{1}{\prod_{a=1}^{3} 2\mathrm{Re}(\psi_2(\mathbf{k}_a))} [2\mathrm{Re}(\psi_3)] \,, \tag{6.4.4a}$$

$$\langle \phi(\mathbf{k}_1)\phi(\mathbf{k}_2)\phi(\mathbf{k}_3)\phi(\mathbf{k}_4) \rangle \sim \frac{1}{\prod_{a=1}^{4} 2\mathrm{Re}(\psi_2(\mathbf{k}_a))} \left[2\mathrm{Re}(\psi_4) + \frac{\mathrm{Re}(\psi_3)\mathrm{Re}(\psi_3)}{\mathrm{Re}(\psi_2)} \right] . \tag{6.4.4b}$$

Here $\mathrm{Re}(\psi_n)$ refers to the real part of the wavefunction coefficient. This is no longer true at loop level, for instance, at one loop we have:

$$\begin{aligned} \langle \phi(\mathbf{k}_1)\phi(\mathbf{k}_2)\phi(\mathbf{k}_3)\phi(\mathbf{k}_4) \rangle \sim & \frac{1}{\prod_{a=1}^{4} 2\mathrm{Re}(\psi_2(\mathbf{k}_a))} \Big[2\mathrm{Re}(\psi_4^{1L}) \\ & + \int \frac{\mathrm{d}^{\mathrm{D}} p}{(2\pi)^{\mathrm{D}}} \frac{1}{2\mathrm{Re}(\psi_2(\mathbf{p}))} \psi_6^{\mathrm{tree}} + \dots \Big] . \end{aligned} \tag{6.4.5}$$

The wavefunction is a more fundamental object than the usual in-in correlators since it contains more information than correlators of fields. It also knows about correlators between fields and their conjugate momentum. If we want to compute the correlator between the fields and its conjugate momentum $\pi(\mathbf{k})$, we can obtain this from the imaginary part of the wavefunction. For example [31]:

$$\langle \pi(\mathbf{k}_1)\phi(\mathbf{k}_2)\phi(\mathbf{k}_3)\rangle \sim \frac{1}{\prod_{a=1}^{3} 2\mathrm{Re}(\psi_2(\mathbf{k}_a))}[2\mathrm{Im}(\psi_3)]\,. \tag{6.4.6}$$

Before we continue, let us review the Feynman rules in the context of cosmological correlators.

6.4.2 Feynman Rules

Perturbatively, given a Feynman diagram, we have the following (schematic) rules:

- To each external line is associated a bulk-to-boundary propagator K. The form of K depends on the spacetime considered and whether the particle propagating is massive. For example:

$$K_\omega = \begin{cases} \mathrm{e}^{i\omega\eta} & \text{(flat space)} \\ (1-i\omega\eta)\mathrm{e}^{i\omega\eta} & \text{(de Sitter massless)} \\ (-\eta)^{3/2}H^{(2)}_{iv}(-\omega\eta) & \text{(de Sitter massive)} \\ -i\partial_\omega \mathrm{e}^{i\omega\eta} & \text{(conformally coupled)} \end{cases}, \tag{6.4.7}$$

  where $H^{(2)}_w$ denotes the Hankel function of the second kind of order $w = iv(\mathrm{D})$, where $v(\mathrm{D}) = \sqrt{\frac{\mathrm{D}^2}{4} - m^2}$.
- To each internal line is associated a bulk-to-bulk propagator G. This propagator can generally be expressed in terms of K's as follows

$$G_\omega(\eta,\eta') = \frac{K^*_\omega(\eta_2)K_\omega(\eta_1)\theta(\eta_1-\eta_2)+(\eta_1\leftrightarrow\eta_2)-K_\omega(\eta_1)K_\omega(\eta_2)}{2\mathrm{Re}(\psi_2)}\,. \tag{6.4.8}$$

- To each vertex are associated a factor F and a time integral from $-\infty < \eta < \eta_0$. In particular, the associated wavefunction will depend only on events happening in its past light cone, which is a signature of causality.

For a more detailed discussion of these rules, see [18, 21, 32, 81].

Example 6.4.1 For example, applying the above rules to the following graph (time is going up) gives

$$\psi_4(\{\omega_k\},\{\mathbf{k}\}) = \quad \begin{array}{c} \\ 1 \quad 2 \quad\quad 3 \quad 4 \\ \quad\quad s \end{array}$$
$$= \int_{-\infty}^{\eta_0} \mathrm{d}\eta_{\mathrm{L}} \int_{-\infty}^{\eta_0} \mathrm{d}\eta_{\mathrm{R}} F_{\mathrm{L}} F_{\mathrm{R}} K_{\omega_1} K_{\omega_2} G_{\omega_s} K_{\omega_3} K_{\omega_4} \,. \tag{6.4.9}$$

Here, "L" and "R" label the left and right vertices, respectively. Note that if the exponentials in the K's converge for negative η, then we have analyticity in the lower half ω planes for the associated wavefunction. We therefore treat $\{\omega_k\}$ as separate variables in which we analytically continue wavefunctions off-shell.

We can say more about the analytic properties of the wavefunction than what was said above. In particular, perturbatively, we can say that generally singularities are localized at (partial) energy poles [32]. It is relatively easy to see that this is at least true for flat-space wavefunctions. We also expect this to hold for massless/conformally coupled scalars in de Sitter: their mode functions can be written as derivatives of a plane wave, which means the wavefunction coefficients can be written as the derivative of the flat-space wavefunction.

Example 6.4.2 When energy going into a subdiagram vanishes, there is a singularity. Thus, given a Feynman diagram, we circle all subdiagrams and equate the energy going in and out. For example, the following are singular loci of the associated diagram

$$\begin{array}{c} 1 \quad 2 \quad\quad 3 \quad 4 \\ \quad\quad s \end{array} \rightsquigarrow \begin{cases} \omega_1 + \omega_2 + \omega_s = 0 \,, \\ \omega_1 + \omega_2 + \omega_3 + \omega_4 = 0 \,, \end{cases} \tag{6.4.10a}$$

$$\begin{array}{c} 1 \quad 2 \quad\quad 3 \quad 4 \end{array} \rightsquigarrow \begin{cases} \omega_1 + \omega_2 + \omega_s + \omega_{s'} = 0 \,, \\ \omega_1 + \omega_2 + \omega_3 + \omega_4 + 2\omega_{s'} = 0 \,. \end{cases} \tag{6.4.10b}$$

Note that all of the above equations have support outside the physical region where $\omega_j > 0$ for all j. This is because there are no physical out states in this formalism, but only in states with positive energy (by convention).

For loop diagrams, the internal energies also depend on the loop momentum $\mathbf{p}$, so we also need to minimize the energy with respect to $\mathbf{p}$. For example, in (6.4.10b), after minimization we obtain (for the flat-space wavefunction of a scalar with mass m):

$$\omega_1 + \omega_2 + \sqrt{s^2 + 4m^2} = 0 \,, \tag{6.4.11}$$

$$\omega_1 + \omega_2 + \omega_3 + \omega_4 + 2m = 0\,. \tag{6.4.12}$$

Here s is the magnitude of the momentum $\mathbf{k}_1 + \mathbf{k}_2$. Notice how the first singularity can be rewritten as $(\omega_1 + \omega_2)^2 - s^2 = 4m^2$.

Moreover, as we saw in [82], the amplitude A is the total energy pole of the wavefunction, i.e.,

$$\psi \sim \frac{A}{\sum_i \omega_i}\,. \tag{6.4.13}$$

At the location of the partial energy poles associated to a subdiagram γ of the original diagram Γ, the wavefunction instead factorizes on lower point amplitudes, i.e.,

$$\psi_\Gamma \sim \frac{A_\gamma \times \tilde{\psi}_{\Gamma\setminus\gamma}}{\big[\sum_{i\in\gamma}\omega_i\big]^{\#_\gamma}}\,. \tag{6.4.14}$$

6.4.3 Bootstrap Constraint 1: Unitarity

Quite like what we have for flat-space scattering amplitudes, we can write the evolution operator [21] as

$$U(\eta_0) = \exp\Big[-i\int_{-\infty}^{\eta_0} \mathrm{d}\eta H_{\rm int}(\eta)\Big]\,, \tag{6.4.15}$$

where $H_{\rm int}$ is the time-dependent interaction Hamiltonian of the system. As a consequence of perturbative unitarity and of being in the Bunch–Davies vacuum, the time evolution operator U is unitary:

$$U^\dagger U = \mathbb{1}\,. \tag{6.4.16}$$

Decomposing $U = \mathbb{1} + \delta U$, unitarity implies that

$$\delta U + \delta U^\dagger = -\delta U \delta U^\dagger\,. \tag{6.4.17}$$

So far, the discussion has been pretty much the same as the standard flat-space one discussed in textbooks. Things differ when we insert bras and kets:

$$\underbrace{\langle n|\delta U|0\rangle}_{\rm (I)} + \underbrace{\langle n|\delta U^\dagger|0\rangle}_{\rm (II)} = -\sum\!\!\!\!\!\!\int_X \underbrace{\langle n|\,\delta U|X\rangle}_{\rm (III)}\langle X|\delta U^\dagger\,|0\rangle\,. \tag{6.4.18}$$

In perturbation theory (when δU is small), it is clear that term (I) corresponds to the wavefunction $\psi_n(\{\omega\})$. What (II) corresponds to is *not* $\psi_n^*(\{\omega\})$, but rather

$\psi_n^*(\{-\omega^*\})$. This observation makes it rather manifest that we *need* analyticity in the energies to define (II) in the first place! Finally, (III) is clearly not $\psi_n(\omega)$ since $|X\rangle$ generally differs from the Bunch–Davies vacuum ($|X\rangle \neq |0\rangle$).

At this stage, let us introduce the *discontinuity operator* in the energy ω:

$$\begin{aligned}\text{Disc}_{\omega_s}\,\psi_n(\omega_1,\ldots,\omega_s,\ldots,\omega_n;\{\mathbf{k}\}) &= \psi_n(\omega_1,\ldots,\omega_s,\ldots,\omega_n;\{\mathbf{k}\})\\ &\quad-\psi_n^*(-\omega_1^*,\ldots,\omega_s,\ldots,-\omega_n^*;\{-\mathbf{k}\})\,.\end{aligned} \tag{6.4.19}$$

Based on this definition, we ask what can we learn about the wavefunction coefficients ϕ_j from this operator?

Example 6.4.3 Let us consider once more the four-point tree-level diagram in (6.4.9). The discontinuity in ω_s is given by

$$\text{Disc}_{\omega_s}\psi_4(\omega_1,\ldots,\omega_4,\omega_s;\{\mathbf{k}\}) = \text{Disc}_{\omega_s}\int_{-\infty}^{\eta_0} d\eta_L d\eta_R F_L F_R K_{\omega_1} K_{\omega_2} G_{\omega_s} K_{\omega_3} K_{\omega_4}\,. \tag{6.4.20}$$

Unitarity ensures that the couplings in H_{int} are real, and so the associated vertex factors satisfy $F(\{\mathbf{k}\}) = F^*(\{-\mathbf{k}\})$, while the fact that we start with the Bunch–Davies vacuum ensures $K(\{\omega\}) = K^*(\{-\omega^*\})$. To see this, remember that the Bunch–Davies vacuum implies $K(\omega) \sim e^{i\omega\eta}$ in the far past, and it is clear that both $K(\omega)$ and $K^*(-\omega^*)$ are the same. It turns out that under some mild assumptions about the equations of motion, this property is preserved under time evolution (see [21] for more details). As an example, for a massless scalar in de Sitter, we have:

$$K^*(-\omega^*) = [(1+i\omega^*\eta)e^{-i\omega^*\eta}]^* = (1-i\omega\eta)e^{i\omega\eta} = K(\omega)\,. \tag{6.4.21}$$

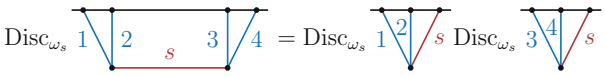
Using this together with (6.4.8) gives

$$\text{Im}G_{\omega_s}(\eta_L,\eta_R) = \frac{1}{2\text{Re}(\psi_2)}\text{Im}(K_{\omega_s}(\eta_L))\text{Im}(K_{\omega_s}(\eta_R))\,, \tag{6.4.22}$$

which gives

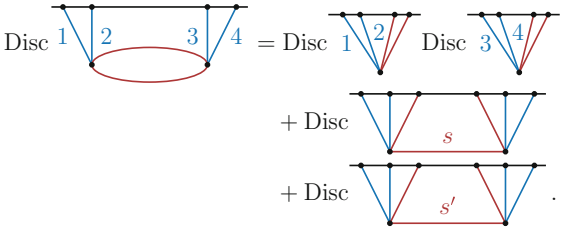
$$\begin{aligned}\text{Disc}_{\omega_s}\psi_4(\omega_1,\ldots,\omega_4,\omega_s;\{\mathbf{k}\}) &= \frac{1}{2\text{Re}(\psi_2)}\int_{-\infty}^{\eta_0} d\eta_L F_L K_{\omega_1} K_{\omega_2}\text{Im}(K_{\omega_s}(\eta_L))\\ &\quad\times\int_{-\infty}^{\eta_0} d\eta_R F_R K_{\omega_3} K_{\omega_4}\text{Im}(K_{\omega_s}(\eta_R))\,.\end{aligned} \tag{6.4.23}$$

Diagrammatically, we have



(6.4.24)

Similar pictures also hold at loop-level. For example, if we take the discontinuity in the loop of the diagram in (6.4.10b), we find [20]



(6.4.25)

The pattern of discontinuities becomes manifest from these two examples: the discontinuity is given by the sum over all ways of cutting the master diagram.

Unitarity already places significant constraints on the bootstrap of the wavefunction coefficients. However, there are additional factors to consider. In the following, we discuss two other constraining properties.

6.4.4 Bootstrap Constraint 2: Manifest Locality

For the rest of this section we will focus mainly on a massless scalar or spin-2 tensor in de Sitter. Since these correspond to scalar curvature perturbations (which seed the fluctuations we observe in the CMB) as well as primordial gravitational waves, these are the most relevant observables in inflationary cosmology.

The bulk-to-boundary propagator for massless scalar in de Sitter looks like [19]

$$K_\omega = (1 - i\omega\eta)\mathrm{e}^{i\omega\eta} \,. \tag{6.4.26}$$

Clearly, K_ω satisfies $\partial_\omega K_\omega|_{\omega=0} = 0$. Therefore, if we assume that all the interactions in $H_{\rm int}$ are built out of fields (and their derivatives) at the same spacetime point, we have obtained a constraint on the wavefunction coefficients which we refer to as the manifest locality test (MLT):

$$\partial_{\omega_e}\psi_n(\{\omega\})|_{\omega_e=0} = 0 \quad \text{for any external energy } \omega_e \,. \tag{6.4.27}$$

Since a massless spin-2 tensor also shares the same bulk-to-boundary propagator, MLT also holds for a massless spin-2 tensor.

6.4.5 Bootstrap Constraint 3: Scale Invariance

Massless scalars and spin-2 tensors in dS_{1+3} scale in the following way:

$$\psi_3 \sim \omega^3 \,. \tag{6.4.28}$$

The reason is the following: in de Sitter, mass and scaling dimension are related by

$$m^2 = \Delta(3-\Delta) \,. \tag{6.4.29}$$

Clearly, if $m^2 = 0$ this implies $\Delta = 0$ or $\Delta = 3$. Consequently a scalar near the future conformal boundary behaves as:

$$\psi_3(\mathbf{x}) \sim \mathcal{O}_{\Delta=0} + \eta^3 \mathcal{O}_{\Delta=3} \,. \tag{6.4.30}$$

As $\eta \to 0$ only the first term survives, and it is easy to see that this implies $\psi_3 \sim \omega^3$ after a Fourier transform.

Since we are in de Sitter, one may wonder if we should impose the full de Sitter isometry, the SO(4,1) group, as a constraint in our bootstrap. However, it can be shown that correlators (and subsequently wavefunction coefficients) for the scalar curvature perturbation are suppressed by the slow roll parameters if they respect the full de Sitter isometry [83].[20] This means that correlators that are relevant for observations in the near future generally break some of the de Sitter isometry. However, since we do observe a power spectrum which scales almost as k^{-3} (as mentioned in Sect. 6.1) we still impose scale invariance as a constraint.

Unitarity, locality, and scale invariance are sufficient to strongly constrain wavefunction coefficients. In the next section, we consider a simple example.

6.4.6 Bootstrapping ψ_3 for a Single Scalar

Let us start by summarizing what we want ψ_3 to satisfy:

- Scale invariance: $\psi_3 \sim \omega^3$.
- Locality: $\partial_\omega \psi_3|_{\omega=0} = 0$.
- Symmetric under permutations of $\omega_{1,2,3}$.

[20] One could also the full de Sitter isometry in the cosmological bootstrap. Initially, the approach is to write down Ward identities from the symmetry, which give rise to differential equations for the wavefunction coefficients [14, 15, 17, 84]. However differential equations have also been found for more general cosmological spacetime [38, 85–89], which leads to the concept of "kinematic flow".

Note that while we have not included unitarity in this list, it will be used in the following section to construct more interesting wavefunction coefficients by "gluing" simpler ones.

The simplest ansatz satisfying these properties is

$$\psi_3^{(p)} = \frac{1}{\omega_T^p} \sum_{m,n\geqslant 0} c_{mn}\omega_T^{3+p-2m-3n} e_2^m e_3^n \text{ where } \left[\begin{array}{l} p \in \mathbb{N} \\ (3+p-2m-3n) \geqslant 0 \\ \omega_T = \omega_1+\omega_2+\omega_3 \\ e_2 = \omega_1\omega_2+\omega_1\omega_3+\omega_2\omega_3 \\ e_3 = \omega_1\omega_2\omega_3 \end{array}\right. . \tag{6.4.31}$$

This ansatz corresponds to the wavefunction coefficients for tree level diagrams. It is written in ω_T, e_2, e_3 which is symmetric in external energies, and its overall scaling is ω^3. Note that p is generally related to the number of derivatives in an interaction.

The only tree-level diagram for ψ_3 is the contact diagram, and so we expect singularities only at total energy poles (as this is the only subdiagram for a contact diagram). Therefore, in our ansatz we only consider the case where $m, n \geqslant 0$ and only allow for total energy poles.

Let us consider some explicit examples.

Example 6.4.4 In de Sitter, let us try to fix (6.4.31) for $p = 0$. We start with the ansatz

$$\psi_3^{(0)} = c_1\omega_T^3 + c_2\omega_T e_2 + c_3 e_3 \,. \tag{6.4.32}$$

Imposing locality gives

$$\partial_{\omega_1}\psi_3^{(0)}|_{\omega_1=0} = 0 \implies c_2 = -3c_1 \text{ and } c_3 = 3c_1 \,, \tag{6.4.33}$$

such that

$$\psi_3^{(0)} = c_1(\omega_T^3 - 3\omega_T e_2 + 3e_3) \,. \tag{6.4.34}$$

This is the result expected from the free theory ($H_{\text{int}} = 0$) after field redefinition $\phi \mapsto \phi + \phi^2$.

We can also modify the ansatz as follows:

$$\psi_3^{(0)} = c_1\omega_T^3 + c_2\omega_T e_2 + c_3 e_3 + \log(-\omega_T\eta_0)(\tilde{c}_1\omega_T^3 + \tilde{c}_2\omega_T e_2 + \tilde{c}_3 e_3) \,. \tag{6.4.35}$$

Repeating the above exercise, we obtain the correct ϕ^3 interaction tree-level wavefunction coefficient.

If we consider $p = 3$, we find only two polynomials which satisfy all the bootstrap requirements:

$$\psi_3^{\text{EFT1}} = \frac{e_3^2}{\omega_T^3}\,, \tag{6.4.36a}$$

$$\psi_3^{\text{EFT2}} = \frac{1}{\omega_T^3}(\omega_T^6 - 3\omega_T^4 e_2 + 11\omega_T^3 e_3 - 4\omega_T^2 e_2^2 - 4\omega_T e_2 e_3 + 12 e_3^2)\,. \tag{6.4.36b}$$

Interestingly they correspond to the wavefunction coefficients computed from $\dot{\phi}^3$ and $\dot{\phi}(\nabla_i\phi)^2$ interactions respectively. Both interactions have exactly three derivatives. This is a general feature: p tell us the number of derivatives on the scalar fields [19].

From this example, we observe that all that we have obtained are tree-level objects. Why is this the case? This is because the ansatz proposed was constructed from simple polynomials. If we aim to derive objects that might emerge from a loop diagram, the ansatz needs to be modified. One loop wavefunctions generally include polylogarithmic or even elliptic functions [32], so we need to add those into the ansatz while taking into account the different possible singularities. Similarly to Feynman integrals [90], in very general cases, it is not even clear what class of functions is needed for the ansatz.

6.4.7 Gluing Procedure

The purpose of this section is to explain how to "glue" tree-level wavefunction coefficients together to create more interesting ones. Below, we illustrate this procedure by gluing ψ_3 with ψ_4. The general approach is as follows.

Starting with (6.4.24), we have (schematically)

$$\begin{aligned}\psi_4(\omega_1,\ldots,\omega_4,\omega_s) - \psi_4^*(-\omega_1,\ldots,-\omega_4,\omega_s) \\ = [\psi_{3,\text{L}}(\omega_1,\omega_2,\omega_s) - \psi_{3,\text{L}}^*(-\omega_1,-\omega_2,\omega_s)] \\ \times [\psi_{3,\text{R}}(\omega_3,\omega_4,\omega_s) - \psi_{3,\text{R}}^*(-\omega_3,-\omega_4,\omega_s)]\,.\end{aligned} \tag{6.4.37}$$

Denoting

$$\omega_1 + \omega_2 + \omega_s = E_{\text{L}} \quad \text{and} \quad \omega_3 + \omega_4 + \omega_s = E_{\text{R}}\,, \tag{6.4.38}$$

as the energy flowing in the left and right vertices respectively and taking the residue at $E_{\text{L}} = 0$ on both sides of (6.4.37) gives

$$\text{Res}_{E_{\text{L}}=0}\psi_4(\omega_1,\ldots,\omega_4,\omega_s)=\psi_{3,\text{R}}(\omega_3,\omega_4,\omega_s)-\psi^*_{3,\text{R}}(-\omega_3,-\omega_4,\omega_s)\,. \tag{6.4.39}$$

Of course, a similar expression holds for the residue at $E_{\text{R}}=0$. Writing the wavefunction coefficients as functions of E_{L} and E_{R}, we can perform the shift $(E_{\text{L}},E_{\text{R}})\mapsto(E_{\text{L}}+z,E_{\text{R}}-z)$ to write

$$B=\oint_\gamma\frac{\text{d}z}{z}\psi_4=\psi_4+\text{Res}_{z=-E_{\text{L}}}\psi_4+\text{Res}_{z=E_{\text{R}}}\psi_4\,. \tag{6.4.40}$$

The contour integral simply picks up the partial energy poles, which are obtained in (6.4.39). By demanding ψ_4 to satisfy the locality test, B can be fixed, and eventually this fixes ψ_4 completely [19].

The takeaway point of this discussion is that, at tree level, unitarity and an ansatz of the form (6.4.31) are sufficient to completely bootstrap the wavefunction coefficients.

Example 6.4.5 As an example of how the gluing procedure works, let us glue two contact ψ_3 from ϕ^3 interaction to obtain an exchange ψ_4. Remember the contact ψ_3 is given by:

$$\psi_3(\omega_1,\omega_2,\omega_3)=\frac{1}{\omega_1+\omega_2+\omega_3}. \tag{6.4.41}$$

First we compute the Disc, which is:

$$\text{Disc}_{\omega_s}\psi_3(\omega_1,\omega_2,\omega_s)=\frac{1}{E_L}-\frac{1}{E_L-2\omega_s}, \tag{6.4.42}$$

where E_L, E_R are defined in (6.4.38). Now we look at the following:

$$\Xi=\frac{1}{2\omega_s}\text{Disc}_{\omega_s}\psi_3(\omega_1,\omega_2,\omega_s)\text{Disc}_{\omega_s}\psi_3(\omega_3,\omega_4,\omega_s), \tag{6.4.43}$$

and shift the energies by $(E_{\text{L}},E_{\text{R}})\mapsto(E_{\text{L}}+z,E_{\text{R}}-z)$. As a result we obtain:

$$\Xi=\frac{1}{2\omega_s}\left[\frac{1}{E_L+z}-\frac{1}{E_L+z-2\omega_s}\right]\left[\frac{1}{E_R-z}-\frac{1}{E_R-z-2\omega_s}\right]. \tag{6.4.44}$$

Clearly the residue at $z=-E_L$ is given by:

$$\frac{1}{(E_R+E_L)(E_R+E_L-2\omega_s)}. \tag{6.4.45}$$

The residue at $z = E_R$ is obtained similarly, and so we have:

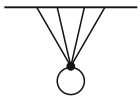
$$\psi_4 = \frac{1}{E_L}\frac{1}{(E_R+E_L)(E_R+E_L-2\omega_s)} + \frac{1}{E_R}\frac{1}{(E_R+E_L)(E_R+E_L-2\omega_s)} + B$$
$$= \frac{1}{\omega_T E_L E_R} + B, \tag{6.4.46}$$

and we notice that $B = 0$ gives us the correct answer for the exchange ψ_4 in flat space.

Generically in de Sitter, Ξ has higher order poles in E_L and E_R, and B would not be zero. See section 6 of [19] for more examples.

6.4.8 Cancellation of Singularities

There is an interesting story about the singularities of in-in correlators versus the singularities of the wavefunction. As an example, consider the one-loop flat-space wavefunction with a single vertex.



(6.4.47)

The wavefunction is given by:

$$\psi_4 \sim \log(\omega_T)\,. \tag{6.4.48}$$

If we try to compute the correlator for the same graph, we find that it is given by:

$$\langle\phi(k_1)\ldots\phi(k_4)\rangle = \int \frac{\mathrm{d}^3 p}{(2\pi)^3}\frac{1}{2p}. \tag{6.4.49}$$

Crucially, no matter how we choose to regularize this integral, we can never obtain a $\log(\omega_T)$ term. If we write down the one-loop in-in correlator in terms of wavefunction coefficients we obtain:

$$\langle\phi(k_1)\ldots\phi(k_4)\rangle_{\text{1-Loop}} \sim \psi_4^{\text{1-Loop}} + \int \frac{\mathrm{d}^3 p}{(2\pi)^3}\frac{1}{2\mathrm{Re}\psi_2(p)}\psi_6^{\text{tree}}. \tag{6.4.50}$$

The integral over the tree-level ψ_6 cancels the logarithmic term from the one-loop ψ_4.

This is an important lesson on the analytic structure of the wavefunction versus the analytic structure of in-in correlators. When we go to loop level, an n-point correlator is no longer only given by ψ_n, but also integrals of higher-

point wavefunctions (with lower loop order). This can result in cancellation of singularities: for instance, we find that any branch point in ω_T from a wavefunction coefficient is never present in an in-in correlator [91]. Interestingly, in flat space and at one loop order, the remaining singularities can be mapped to singularities from an amplitude with the same Feynman graph [92]. It would be interesting to better understand this story, particularly in the context of de Sitter spacetime.[21]

References

1. S. Dodelson, Coherent phase argument for inflation. AIP Conf. Proc. **689**, 184 (2003) [hep-ph/0309057]
2. Planck Collaboration, R. Adam et al., Planck 2015 results. I. Overview of products and scientific results. Astron. Astrophys. **594**, A1 (2016) [1502.01582]
3. Planck Collaboration, Y. Akrami et al., Planck 2018 results. X. Constraints on inflation. Astron. Astrophys. **641**, A10 (2020) [1807.06211]
4. Planck Collaboration, Y. Akrami et al., Planck 2018 results. IX. Constraints on primordial non-Gaussianity. Astron. Astrophys. **641**, A9 (2020) [1905.05697]
5. Planck Collaboration, N. Aghanim et al., Planck 2018 results. VI. Cosmological parameters. Astron. Astrophys. **641**, A6 (2020) [1807.06209]
6. G. Cabass, E. Pajer, F. Schmidt, How Gaussian can our Universe be?, J. Cosmol. Astropart. Phys. **01**, 003 (2017) [1612.00033]
7. A. Einstein, The foundation of the general theory of relativity. Ann. Phys. **49**, 769 (1916)
8. A. Achúcarro et al., Inflation: Theory and Observations. 2203.08128
9. G.L. Pimentel, Inflationary consistency conditions from a wavefunctional perspective. J. High Energy Phys. **02**, 124 (2014) [1309.1793]
10. J.M. Maldacena, Non-Gaussian features of primordial fluctuations in single field inflationary models. J. High Energy Phys. **05**, 013 (2003) [astro-ph/0210603]
11. D. Anninos, T. Anous, D.Z. Freedman, G. Konstantinidis, Late-time structure of the Bunch-Davies De Sitter wavefunction. J. Cosmol. Astropart. Phys. **11**, 048 (2015) [1406.5490]
12. A. Ghosh, N. Kundu, S. Raju, S.P. Trivedi, Conformal invariance and the four point scalar correlator in slow-roll inflation. J. High Energy Phys. **07**, 011 (2014) [1401.1426]
13. N. Arkani-Hamed, P. Benincasa, A. Postnikov, Cosmological Polytopes and the Wavefunction of the Universe. 1709.02813.
14. N. Arkani-Hamed, D. Baumann, H. Lee, G.L. Pimentel, The cosmological bootstrap: inflationary correlators from symmetries and singularities. J. High Energy Phys. **04**, 105 (2020) [1811.00024]
15. D. Baumann, C. Duaso Pueyo, A. Joyce, H. Lee, G.L. Pimentel, The cosmological bootstrap: weight-shifting operators and scalar seeds. J. High Energy Phys. **12**, 204 (2020) [1910.14051]
16. A. Hillman, Symbol Recursion for the dS Wave Function, 1912.09450
17. D. Baumann, C. Duaso Pueyo, A. Joyce, H. Lee, G.L. Pimentel, The cosmological bootstrap: spinning correlators from symmetries and factorization. SciPost Phys. **11**, 071 (2021) [2005.04234]
18. H. Goodhew, S. Jazayeri, E. Pajer, The cosmological optical theorem. J. Cosmol. Astropart. Phys. **04**, 021 (2021) [2009.02898]

[21] The flat space result was obtained by considering the spectral representation of the bulk-to-bulk propagator, which give rise to an integral representation of the wavefunction coefficients. Naturally one could try developing similar integral representations [93], spectral representations or dispersive relations [56, 94–97] in de Sitter.

19. S. Jazayeri, E. Pajer, D. Stefanyszyn, From locality and unitarity to cosmological correlators. J. High Energy Phys. **10**, 065 (2021) [2103.08649]
20. S. Melville, E. Pajer, Cosmological cutting rules. J. High Energy Phys. **05**, 249 (2021) [2103.09832]
21. H. Goodhew, S. Jazayeri, M.H.G. Lee, E. Pajer, Cutting cosmological correlators. J. Cosmol. Astropart. Phys. **08**, 003 (2021) [2104.06587]
22. J. Bonifacio, E. Pajer, D.-G. Wang, From amplitudes to contact cosmological correlators. J. High Energy Phys. **10**, 001 (2021) [2106.15468]
23. G. Cabass, D. Stefanyszyn, J. Supeł, A. Thavanesan, On graviton non-Gaussianities in the effective field theory of inflation. J. High Energy Phys. **10**, 154 (2022) [2209.00677]
24. D. Stefanyszyn, X. Tong, Y. Zhu, *Cosmological correlators through the looking glass: reality, parity, and factorisation*. J. High Energy Phys. **05**, 196 (2024) [2309.07769]
25. G. Cabass, E. Pajer, D. Stefanyszyn, J. Supeł, *Bootstrapping large graviton non-Gaussianities*. J. High Energy Phys. **05**, 077 (2022) [2109.10189]
26. A. Hillman, E. Pajer, A differential representation of cosmological wavefunctions. J. High Energy Phys. **04**, 012 (2022) [2112.01619]
27. G. Cabass, S. Jazayeri, E. Pajer, D. Stefanyszyn, Parity violation in the scalar trispectrum: no-go theorems and yes-go examples. J. High Energy Phys. **02**, 021 (2023) [2210.02907]
28. J. Bonifacio, H. Goodhew, A. Joyce, E. Pajer, D. Stefanyszyn, The graviton four-point function in de Sitter space. J. High Energy Phys. **06**, 212 (2023) [2212.07370]
29. G.L. Pimentel, D.-G. Wang, Boostless cosmological collider bootstrap. J. High Energy Phys. **10**, 177 (2022) [2205.00013]
30. D.-G. Wang, G.L. Pimentel, A. Achúcarro, Bootstrapping multi-field inflation: non-Gaussianities from light scalars revisited. J. Cosmol. Astropart. Phys. **05**, 043 (2023) [2212.14035]
31. S. Céspedes, A.-C. Davis, S. Melville, On the time evolution of cosmological correlators. J. High Energy Phys. **02**, 012 (2021) [2009.07874]
32. S.A. Salcedo, M.H.G. Lee, S. Melville, E. Pajer, The analytic wavefunction. J. High Energy Phys. **06**, 020 (2023) [2212.08009]
33. M.H.G. Lee, C. McCulloch, E. Pajer, Leading loops in cosmological correlators. J. High Energy Phys. **11**, 038 (2023) [2305.11228]
34. C. Duaso Pueyo, E. Pajer, A cosmological bootstrap for resonant non-Gaussianity. J. High Energy Phys. **03**, 098 (2024) [2311.01395]
35. D. Ghosh, E. Pajer, F. Ullah, Cosmological cutting rules for Bogoliubov initial states. SciPost Phys. **18**, 005 (2025) [2407.06258]
36. D. Stefanyszyn, X. Tong, Y. Zhu, There and back again: mapping and factorising cosmological observables. 2406.00099
37. H. Goodhew, A. Thavanesan, A.C. Wall, The Cosmological CPT Theorem. 2408.17406
38. P. Benincasa, G. Brunello, M.K. Mandal, P. Mastrolia, F. Vazão, On one-loop corrections to the Bunch-Davies wavefunction of the universe. 2408.16386
39. S. Céspedes, A.-C. Davis, D.-G. Wang, On the IR divergences in de Sitter space: loops, resummation and the semi-classical wavefunction. J. High Energy Phys. **04**, 004 (2024) [2311.17990]
40. S. Jazayeri, S. Renaux-Petel, Cosmological bootstrap in slow motion. J. High Energy Phys. **12**, 137 (2022) [2205.10340]
41. P. Creminelli, S. Renaux-Petel, G. Tambalo, V. Yingcharoenrat, Non-perturbative wavefunction of the universe in inflation with (resonant) features. J. High Energy Phys., 010 **03** (2024) [2401.10212]
42. G. Date, S.J. Hoque, Gravitational waves from compact sources in a de Sitter background. Phys. Rev. D **94**, 064039 (2016) [1510.07856]
43. J. de Boer, V. Jejjala, D. Minic, Alpha-states in de Sitter space. Phys. Rev. D **71**, 044013 (2005) [hep-th/0406217]
44. M.B. Einhorn, F. Larsen, Squeezed states in the de Sitter vacuum. Phys. Rev. D **68**, 064002 (2003) [hep-th/0305056]

45. D. Ghosh, A.H. Singh, F. Ullah, Probing the initial state of inflation: analytical structure of cosmological correlators. J. Cosmol. Astropart. Phys. **04**, 007 (2023) [2207.06430]
46. A.J. Chopping, C. Sleight, M. Taronna, Cosmological correlators for Bogoliubov initial states. J. High Energy Phys. **09**, 152 (2024) [2407.16652]
47. X. Chen, Y. Wang, Z.-Z. Xianyu, Schwinger-Keldysh Diagrammatics for Primordial Perturbations. J. Cosmol. Astropart. Phys. **12**, 006 (2017) [1703.10166]
48. Y. Donath, E. Pajer, The in-out formalism for in-in correlators. J. High Energy Phys. **07**, 064 (2024) [2402.05999]
49. A. Kamenev, *Field Theory of Non-equilibrium Systems* (University Press, Cambridge, 2011)
50. Y. Ema, K. Mukaida, Cutting rule for in-in correlators and cosmological collider. J. High Energy Phys. **12**, 194 (2024) [2409.07521]
51. M.J.G. Veltman, *Diagrammatica: The Path to Feynman Rules*, vol. 4 (Cambridge University Press, Cambridge, 2012)
52. F.M. Haehl, M. Rangamani, Records from the S-Matrix Marathon: Schwinger-Keldysh Formalism (10, 2024). 2410.10602
53. S. Caron-Huot, M. Giroux, H.S. Hannesdottir, S. Mizera, C. Pasiecznik, Records from the S-Matrix Marathon: Asymptotic Observables. [2502.13021]
54. D. Marolf, I.A. Morrison, M. Srednicki, Perturbative S-matrix for massive scalar fields in global de Sitter space. Class. Quant. Grav. **30**, 155023 (2013) [1209.6039]
55. S. Melville, G.L. Pimentel, A de Sitter *S*-matrix for the masses. 2309.07092
56. S. Melville, G.L. Pimentel, A de Sitter S-matrix from amputated cosmological correlators. 2404.05712
57. D. Baumann, D. Green, H. Lee, R.A. Porto, Signs of analyticity in single-field inflation. Phys. Rev. D, 023523 **93** (2016) [1502.07304]
58. T. Grall, S. Melville, Inflation in motion: unitarity constraints in effective field theories with (spontaneously) broken Lorentz symmetry. J. Cosmol. Astropart. Phys. **09**, 017 (2020) [2005.02366]
59. T. Grall, S. Melville, Positivity bounds without boosts: new constraints on low energy effective field theories from the UV. Phys. Rev. D **105**, L121301 (2022) [2102.05683]
60. E. Pajer, D. Stefanyszyn, J. Supeł, The Boostless Bootstrap: amplitudes without Lorentz boosts. J. High Energy Phys. **12**, 198 (2020) [2007.00027]
61. L. Hui, I. Kourkoulou, A. Nicolis, A. Podo, S. Zhou, S-matrix positivity without Lorentz invariance: a case study. J. High Energy Phys. **04**, 145 (2024) [2312.08440]
62. P. Creminelli, O. Janssen, L. Senatore, Positivity bounds on effective field theories with spontaneously broken Lorentz invariance. J. High Energy Phys. **09**, 201 (2022) [2207.14224]
63. P. Creminelli, M. Delladio, O. Janssen, A. Longo, L. Senatore, Non-analyticity of the S-matrix with spontaneously broken Lorentz invariance. J. High Energy Phys. **06**, 201 (2024) [2312.08441]
64. P. Creminelli, O. Janssen, B. Salehian, L. Senatore, Positivity bounds on electromagnetic properties of media. J. High Energy Phys. **08**, 066 (2024) [2405.09614]
65. N. Arkani-Hamed, J. Maldacena, Cosmological collider physics. 1503.08043
66. T. Noumi, M. Yamaguchi, D. Yokoyama, Effective field theory approach to quasi-single field inflation and effects of heavy fields. J. High Energy Phys. **06**, 051 (2013) [1211.1624]
67. P. Creminelli, S. Kumar, B. Salehian, L. Santoni, Dissipative inflation via scalar production. J. Cosmol. Astropart. Phys. **2023**, 076 (2023)
68. A.A. Starobinsky, Stochastic de sitter (inflationary) stage in the early universe. Lect. Notes Phys. **246**, 107 (1986)
69. A.A. Starobinsky, J. Yokoyama, Equilibrium state of a selfinteracting scalar field in the De Sitter background. Phys. Rev. D **50**, 6357 (1994) [astro-ph/9407016]
70. S.A. Salcedo, T. Colas, E. Pajer, The open effective field theory of inflation. J. High Energy Phys. **10**, 248 (2024) [2404.15416]
71. T. Colas, Open effective field theories for cosmology, in *58th Rencontres de Moriond on Cosmology*, vol. 5 (2024) 2405.09639

72. D. Lopez Nacir, R.A. Porto, L. Senatore, M. Zaldarriaga, Dissipative effects in the effective field theory of inflation. J. High Energy Phys. **01**, 075 (2012) [1109.4192]
73. M. Hongo, S. Kim, T. Noumi, A. Ota, Effective field theory of time-translational symmetry breaking in nonequilibrium open system. J. High Energy Phys. **02**, 131 (2019) [1805.06240]
74. M. Crossley, P. Glorioso, H. Liu, Effective field theory of dissipative fluids. J. High Energy Phys. **09**, 095 (2017) [1511.03646]
75. H. Liu, P. Glorioso, Lectures on non-equilibrium effective field theories and fluctuating hydrodynamics. Proc. Sci. **TASI2017**, 008 (2018) [1805.09331]
76. A. Adams, N. Arkani-Hamed, S. Dubovsky, A. Nicolis, R. Rattazzi, Causality, analyticity and an IR obstruction to UV completion. J. High Energy Phys. **10**, 014 (2006) [hep-th/0602178]
77. B. Finelli, G. Goon, E. Pajer, L. Santoni, The effective theory of shift-symmetric cosmologies. J. Cosmol. Astropart. Phys. **05**, 060 (2018) [1802.01580]
78. M. Hongo, S. Kim, T. Noumi, A. Ota, Effective Lagrangian for Nambu-Goldstone modes in nonequilibrium open systems. Phys. Rev. D **103**, 056020 (2021) [1907.08609]
79. C.O. Akyuz, G. Goon, R. Penco, The Schwinger-Keldysh Coset construction. 2306.17232
80. E. Pajer, Building a boostless bootstrap for the bispectrum. J. Cosmol. Astropart. Phys. **01**, 023 (2021) [2010.12818]
81. D. Baumann, A. Joyce, Lecture notes on cosmological correlators. (2014)
82. P. Benincasa, M. Giroux, H.S. Hannesdottir, S. Mizera, C. Pasiecznik, F. Vazão, Records from the S-Matrix Marathon: observables in expanding universes (2024). 2409.14947
83. D. Green, E. Pajer, On the symmetries of cosmological perturbations. J. Cosmol. Astropart. Phys. **09**, 032 (2020) [2004.09587]
84. S. Aoki, L. Pinol, F. Sano, M. Yamaguchi, Y. Zhu, Cosmological correlators with double massive exchanges: bootstrap equation and phenomenology. J. High Energy Phys. **09**, 176 (2024) [2404.09547]
85. N. Arkani-Hamed, D. Baumann, A. Hillman, A. Joyce, H. Lee, G.L. Pimentel, Differential equations for cosmological correlators. 2312.05303
86. N. Arkani-Hamed, D. Baumann, A. Hillman, A. Joyce, H. Lee, G.L. Pimentel, Kinematic flow and the emergence of time. 2312.05300
87. T.W. Grimm, A. Hoefnagels, M. van Vliet, Structure and complexity of cosmological correlators. Phys. Rev. D **110**, 123531 (2024) [2404.03716]
88. S. He, X. Jiang, J. Liu, Q. Yang, Y.-Q. Zhang, Differential equations and recursive solutions for cosmological amplitudes. J. High Energy Phys. **01**, 001 (2025) [2407.17715]
89. T.W. Grimm, A. Hoefnagels, Reductions of GKZ Systems and Applications to Cosmological Correlators. 2409.13815
90. J.L. Bourjaily et al., Functions beyond multiple polylogarithms for precision collider physics, in *Snowmass 2021* (2022). 2203.07088
91. S. Agui Salcedo, S. Melville, The cosmological tree theorem. J. High Energy Phys. **12**, 076 (2023) [2308.00680]
92. M.H.G. Lee, From amplitudes to analytic wavefunctions. J. High Energy Phys. **03**, 058 (2024) [2310.01525]
93. C. Chowdhury, A. Lipstein, J. Mei, I. Sachs, P. Vanhove, The subtle simplicity of cosmological correlators. 2312.13803
94. D. Meltzer, The inflationary wavefunction from analyticity and factorization. J. Cosmol. Astropart. Phys. **12**, 018 (2021) [2107.10266]
95. C. Sleight, M. Taronna, From AdS to dS exchanges: spectral representation, Mellin amplitudes, and crossing. Phys. Rev. D **104**, L081902 (2021) [2007.09993]
96. D. Werth, Spectral representation of cosmological correlators. J. High Energy Phys. **12**, 017 (2024) [2409.02072]
97. H. Liu, Z. Qin, Z.-Z. Xianyu, Dispersive bootstrap of massive inflation correlators. 2407.12299

Gravitational Physics from Scattering Amplitudes

7

Miguel Correia, Holmfridur Sigridar Hannesdottir, Giulia Isabella, Anna M. Wolz, and Zihan Zhou

Abstract

These lecture notes explain how classical gravitational physics emerges from scattering amplitudes. We emphasize the role of different kinematic regimes in probing various aspects of bound and unbound problems, as illustrated by the Hydrogen atom example. Classical predictions of General Relativity, such as the Shapiro time delay and perihelion precession, emerge from these considerations. We also explain a number of recent approaches to probing black hole physics from the perspective of amplitudes, including applications of worldline effective field theory in astrophysics, predictions of gravitational waveforms, and the hierarchical three-body problem.

M. Correia
Department of Physics, McGill University, Montréal, QC, Canada
e-mail: miguel.ribeirocorreia@mcgill.ca

H. S. Hannesdottir
Institute for Advanced Study, Princeton, NJ, USA
e-mail: hofie@ias.edu

G. Isabella (✉) · A. M. Wolz
Department of Physics and Astronomy, Mani L. Bhaumik Institute for Theoretical Physics, University of California Los Angeles, Los Angeles, CA, USA
e-mail: giuliaisabella@physics.ucla.edu; awolz@g.ucla.edu

Z. Zhou
Department of Physics, Princeton University, Princeton, NJ, USA
e-mail: zz0962@princeton.edu

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025

N. Arkani-Hamed et al. (eds.), *Records from the S-Matrix Marathon*,
Lecture Notes in Physics 1041, https://doi.org/10.1007/978-3-031-90352-6_7

7.1 Introduction

These lectures will focus on how to use the classical limit of quantum field theory (QFT) to compute observables, in particular in electromagnetism and general relativity (GR). In other contributions to [1], we have learned about the beautiful analytic structure of the S-matrix in QFT. Here, we will examine to what extent this structure survives in the classical limit, and what insight it offers for classical physics.

The question of how to take the classical limit of QFT is very old, but it is particularly relevant today. Since gravitational waves were detected at LIGO [2], the application of scattering-amplitude techniques to the computation of gravitational observables has become an active area of research. Indeed, people have realized that perturbative techniques in QFT, developed over decades in the context of making predictions for the Large Hadron Collider at CERN, can be directly applied to compute quantities that are otherwise hard to obtain directly from General Relativity (GR).

Interestingly, many physical observables in GR (scattering angle, perihelion shifts, etc.) must be computed non-perturbatively, and thus provide an interesting case study for resummation of perturbative expansions within QFT. A simple dimensional-analysis argument shows why we need to go beyond perturbation theory: since the gravitational constant $G_N = 1/M_{\rm Pl}^2$ is dimensionful (where $M_{\rm Pl}$ is the Planck mass), it must be multiplied by some kinematic quantity, which can be taken to be the center-of-mass energy, in order to become a dimensionless perturbative coupling. The relevant coupling is therefore the center-of-mass energy squared in units of the Planck mass: $E_{\rm cm}^2/M_{\rm Pl}^2$. Since the Planck mass is of order $M_{\rm Pl} \sim 10^{-8}$ kg, this coupling constant becomes $\mathcal{O}(1)$ even for the scattering of mosquitoes. Extracting classical observables thus requires a sufficiently thorough understanding of scattering-amplitude techniques to make *all-orders* statements. In these lectures, we will review some of the things that are known to hold at all orders in perturbation theory.

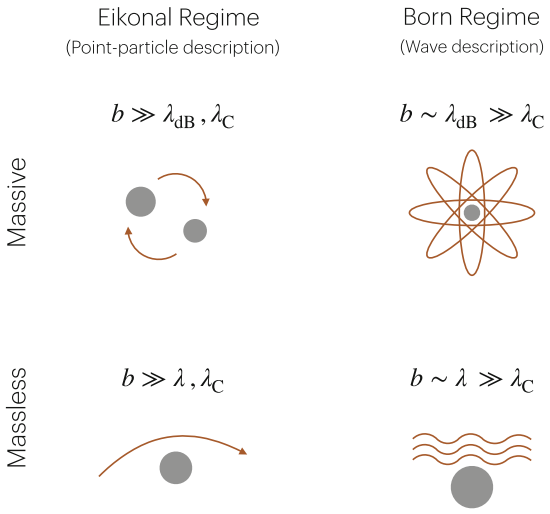
Before going on, let us clarify what we mean by the classical limit. We have the following kinematic scales in our problem (dynamics will come later):

- Compton wavelength, $\lambda_{\rm C} = \frac{\hbar}{m}$, where m is the lightest mass participating in the scattering event. This scale is associated with particle production and QFT effects (precisely the ones we want to eliminate in a classical setup).
- De Broglie wavelength (for massive particles), $\lambda_{\rm dB} = \frac{\hbar}{|p|}$, associated with quantum-mechanical effects, where the wave nature of the particle becomes important.
- The impact parameter b, which is the typical separation length between the scattered bodies.

In order to get rid of all the particle production/QFT effects, we will always consider

$$b \gg \lambda_C \,. \tag{7.1.1}$$

We can then distinguish between four interesting regimes:



(7.1.2)

The first one involves a massive body in a point-particle approximation (top left). This corresponds to the regime $b \gg \lambda_{dB}, \lambda_C$. Similarly, we can remain in the point-particle setup, but consider a massless particle traveling along a null geodesic (bottom left). The regime relevant to this event is $b \gg \lambda, \lambda_C$, where λ is the wavelength of the massless particle. There are also wave counterparts to both regimes. If the de Broglie wavelength of a massive particle is comparable to the size of the experiment (top right), we have $b \sim \lambda_{dB} \gg \lambda_C$. Notice that this inequality implies a non-relativistic setup where $|p| \ll m$. We will discuss this regime at length later in the context of the non-relativistic hydrogen atom. Likewise, if the wavelength of a massless particle is of a size comparable to the impact parameter (bottom right), we are in the $b \sim \lambda \gg \lambda_C$ regime. In this case, this is a probe limit, where the frequency of the wave is much smaller than any of the masses present in the problem.

For our purpose of describing gravitational scattering, the most important situations will be the first and the last: either massive scattering in a point-particle approximation, or a massless wave scattering in the probe limit. Indeed, the actual observable we measure in gravitational wave physics is a signal that looks something like Fig. 7.1. The signal can be divided into three different regions in time. The first one is the inspiral phase, where two black holes or other heavy objects orbit around each other and slowly coalesce. This phase corresponds to the massive point-particle regime. The middle phase is the merger. It is a strongly

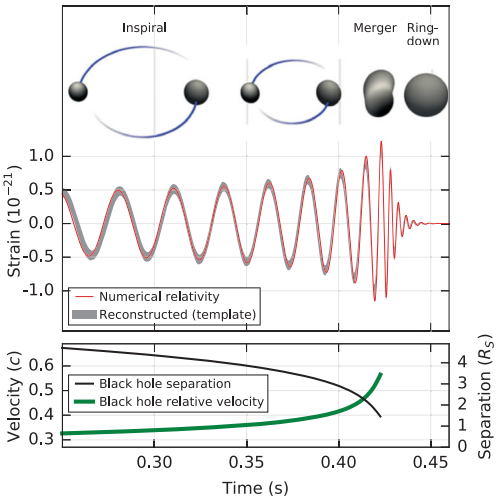


Fig. 7.1 The gravitational-wave signal from the GW150914 event, illustrating the three phases of a binary black hole coalescence. Reprinted under CC-BY-3.0 license from [2]. © 2016, The Author(s)

coupled regime where the two black holes collapse, and numerical GR techniques are needed. The final phase is the ringdown, where the system settles into a single black hole and can be thought of as a perturbation over the Schwarzschild metric. It corresponds to the massless wave regime explained above.

The outline of this chapter is as follows. First, we will explore in detail the case of the non-relativistic hydrogen atom. It turns out that QED in the non-relativistic limit is a wonderful case study, which despite being computationally much simpler, allows us to introduce all the main ingredients that will be needed later. Then, once we have a solid understanding in QED, we will apply an analogous setup to gravity. Finally, the latter sections focus on absorption effects, radiation, and the three-body problem, hopefully offering a more complete and broad view on this active and fascinating subject.

7.2 Feynman Diagrams, Hydrogen Atom, and Classical Limit

Giulia Isabella

Disclaimer: The first two sections of this chapter are the result of some work and study of the literature over the last months. This is of course a huge and active subject with thousands of papers, and the content of these sections will be a personal biased view of the authors, but hopefully a coherent picture of the subject. The references that we will mostly use are:

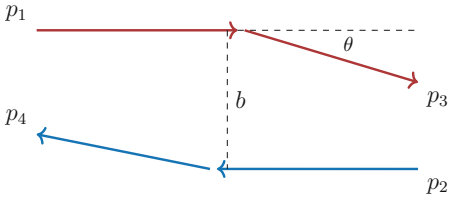
- "Quasipotential equation corresponding to the relativistic eikonal approximation" by Todorov [3].
- "Analyticity in the complex angular momentum plane of the coulomb scattering amplitude" by Singh [4].

- "From scattering amplitudes to classical potentials in the post-Minkowskian expansion" by Cheung et al. [5].
- "Post-Minkowskian effective field theory for conservative binary dynamics" by Kälin and Porto [6].
- "Scalar QED as a toy model for higher-order effects in classical gravitational scattering" by Bern et al. [7].
- "Classical vs quantum eikonal scattering and its causal structure" by Bellazzini et al. [8].
- "The Born regime of gravitational amplitudes" by Correia and Isabella [9].

7.2.1 Classical Observables for the Hydrogen Atom

In this section, we will model the hydrogen atom from scattering amplitudes and study its various kinematic limits. Our first goal is to show how to compute observables in the limit of the top left corner of (7.1.2), which in momentum space corresponds to taking the limit of transferred momentum to be smaller than the masses and momenta for any individual particle. We could think of this regime as throwing some macroscopic charged balls that pass next to each other, where the electromagnetic attraction/repulsion will barely affect their trajectories.

The $2 \to 2$ scattering configuration that we have in mind in impact parameter space is as follows:



(7.2.1)

Let us call the mass, momentum and charge of the particles m_i, p_i and Q_i respectively, for $i = 1, 2$. The kinematics can be described in terms of the scattering angle θ, the center-of-mass energy squared $s = \mathbf{p}^2 = (p_1 + p_2)^2 > 0$, and the momentum transfer squared $t = -\mathbf{q}^2 = (p_1 + p_3)^2 < 0$. They satisfy the constraint

$$\cos\theta = 1 - t/2\mathbf{p}^2 \,. \tag{7.2.2}$$

We use mostly-minus metric signature. Bold faced quantities denote 3-vectors, such that $p_i^2 = (p_i^0)^2 - \mathbf{p}_i^2 = m_i^2$.

As mentioned earlier, the point particle regime we are interested in is given by $|\mathbf{q}| \ll m_i, |\mathbf{p}|$. Additionally, we are going to study this scattering process in the simplest possible setup: non-relativistically, in the probe limit ($m_1 \gg m_2$), and ignoring spin effects by considering scalar QED for simplicity.

Let us start by drawing and computing diagrams for this process. At tree level, we simply have the contribution from the photon exchange diagram, which in this limit evaluates to

$$\mathcal{M}_0 = \quad = -4m_1m_2\frac{Q_1Q_2}{\mathbf{q}^2}\,. \tag{7.2.3}$$

Recall that in this notation $\mathbf{q}$ is the (spatial) momentum transfer. We are interested in the amplitude in the b space, where the impact parameter b is the separation between the two bodies in the plane transverse to the scattering as illustrated in (7.2.1). It turns out that at small momentum transfer, the vector $\mathbf{q}$ is essentially two-dimensional. So, the next step is to simply (2D) Fourier transform the above result to the b space. The result is

$$\mathrm{FT}[\mathcal{M}_0] = -Q_1Q_2\log b/b_{\mathrm{IR}}\,. \tag{7.2.4}$$

which is IR divergent and we have regularized it with a sharp IR cutoff b_{IR}.

We will continue the computation at the one-loop level. It turns out that in this limit the only diagram that contributes to leading order is the planar box. Its contribution can be readily computed and gives

$$\mathcal{M}_1 = \quad = i\frac{Q_1^2Q_2^2m_1m_2}{\pi\mathbf{q}^2\sqrt{2E/m_2}}\log\mathbf{q}^2\,. \tag{7.2.5}$$

Here, $E = |\mathbf{p}|^2/2m_2$ is the non-relativistic kinetic energy of the probe particle. Its Fourier transform is

$$\mathrm{FT}[\mathcal{M}_1] = i\frac{(Q_1Q_2\log b/b_{\mathrm{IR}})^2}{\sqrt{E/m_2}}\,. \tag{7.2.6}$$

This pattern continues. If you perform a two-loop computation, you get a term proportional to $(\log b)^3$ from the double box, etc. Taking care of all the factors of 2 and overall normalizations (we will fix them later), it is easy to see that the computation using the diagrams

$$\quad + \quad + \quad + \quad + \ldots \tag{7.2.7}$$

exponentiates to

$$S(E,b) = \mathrm{e}^{2i\delta(E,b)} = \mathrm{e}^{-i\frac{2Q_1Q_2}{\sqrt{2E/m_2}}\log b/b_{\mathrm{IR}}} = \mathrm{e}^{-2i\alpha_e\log b/b_{\mathrm{IR}}}\,, \tag{7.2.8}$$

in the impact parameter space. Note that we included the disconnected piece, which gives the 1 in a small-α_e expansion of the exponential (the same 1 as the matrix element of $\mathbb{1}$ in $S = \mathbb{1} + iT$). Here

$$\alpha_e = \frac{Q_1 Q_2}{\sqrt{2E/m_2}}\,, \tag{7.2.9}$$

is the effective QED coupling. We will refer to $\delta(E, b)$ as the *phase shift*. Notice that restoring the $\hbar$ gives you the phase shift $\delta(E, b) = -\frac{1}{\hbar}\alpha_e \log b/b_{\rm IR}$ so the loop diagrams are in fact more classical, i.e., more superleading to the tree diagrams as $\hbar \to 0$. These contributions are usually referred to as superclassical. Later on, we will see that an analogous exponentiation takes place for gravity: it is basically a resummation of the Newtonian potential.

We are now interested in extracting some classical observables from this expression. We know that the amplitude is a function of the transferred momentum $\mathbf{q}$ and the energy of the probe particle E. We already saw that the transferred momentum is conjugate to the impact parameter b. Since time translation acts on the amplitude as a phase e^{iHT}, where H is the Hamiltonian, it is easy to see that the energy is conjugate to some time delay. If we assume that we scatter wave packets sharply peaked at a given energy, then the inverse transform (from b to $\mathbf{q}$) will look like[1]

$$\int \mathrm{d}E \int \mathrm{d}^2\mathbf{b}\, e^{2i\delta(E,b)+i\mathbf{q}\cdot\mathbf{b}+iTE}\,, \tag{7.2.10}$$

where $b = |\mathbf{b}|$. This is a highly-oscillatory integral. It is dominated by two saddle points. One appearing when integrating over E, which leads to

$$T = \frac{2\partial \mathrm{Re}\,\delta(E, b)}{\partial E} = \frac{-Q_1 Q_2}{\sqrt{2} m_2 (E/m_2)^{3/2}} \log b\,, \tag{7.2.11}$$

and a second one emerging from the $\mathbf{b}$ integration

$$\theta = \frac{2}{|\mathbf{p}|}\frac{\partial \mathrm{Re}\delta(E, b)}{\partial b} = -\frac{Q_1 Q_2}{Eb}\,, \tag{7.2.12}$$

where we used $\theta \sim \frac{|\mathbf{q}|}{|\mathbf{p}|}$. Hence, this calculation recovers the classical scattering angle and time delay!

Notice that in order to develop a saddle point, we need a large oscillating phase that only happens when we scatter objects with macroscopic charges. In Sect. 7.3, we will do a similar analysis and discuss what regime gives measurable observables

[1] Note that we are using a slight abuse of notation between 2D and 3D vectors. For simplicity we can think of $\mathbf{b}$ as a 3D vector with non-zero components only in the xy plane.

for gravity (the same exercise for gravity leads to the Shapiro time delay and the leading scattering angle R_s/b).

7.2.2 Large Angular Momentum Limit of Partial Waves

Up to this point, everything worked rather nicely at the leading order, but we have been somewhat heuristic. Let us try to do this computation more carefully.

When the impact parameter b is very large, the classical angular momentum J also becomes large. Therefore, it is easy to imagine that we can access this regime by projecting the amplitude in partial waves and taking the large J limit.

We can carry this construction out very explicitly for the hydrogen atom. It turns out that the resummation can be performed directly in the $\mathbf{q}$ space for this simple case. The result is

$$\mathcal{M} \sim \frac{Q_1 Q_2}{t} \frac{\Gamma(1 - i\alpha_e)}{\Gamma(1 + i\alpha_e)} \left(\frac{-t}{\mu^2}\right)^{i\alpha_e} , \tag{7.2.13}$$

where μ is the scale from dimensional regularization that cuts off the IR divergences. Notice the appearance of an infinite series of poles coming from the Gamma function in the numerator. Of course, these are associated with the energy levels of the hydrogen atom. One might wonder why we are observing energy levels, which are quantum mechanical properties of the wave regime, in this classical point-particle limit. This is a special feature of the hydrogen atom: in momentum space (and non-relativistic limit), the amplitude is the same for both the wave and point particle regime. This is an accident that does not hold in gravity.

It is more standard to see how this plays out in the J space. We can then compute the partial waves which (up to an overall IR divergent phase) are

$$S_J(E) = \frac{\Gamma(1 + J - i\alpha_e)}{\Gamma(1 + J + i\alpha_e)} , \tag{7.2.14}$$

where once again we have included the disconnected piece. Let us now take the large J limit of this expression. Using the Stirling approximation, it is easy to see that the leading contribution is indeed

$$\mathrm{e}^{-2i\alpha_e \log J} , \tag{7.2.15}$$

where $J = |\mathbf{p}|b = \sqrt{2m_2 E}b$ is the classical angular momentum and the IR divergence is just the overall phase. This expression thus matches the naive Fourier transform that we did earlier! It is clearly the right direction.

Of course, in general, we will not be able to access the full partial-waves solution. Therefore, let us study this limit directly at the level of the transform.

For completeness, let us do this derivation with spinning external states. Partial waves are representations of the little group in the center-of-mass frame that is

SO(3). They can be computed as

$$\mathcal{M}_J{}^{\lambda_3\lambda_4}_{\lambda_1\lambda_2}(s) = \mathcal{N}^{-1}\int_{-1}^{1} \mathrm{d}\cos\theta\; d^J_{\lambda_{12}\lambda_{34}}(\theta)\,\mathcal{M}^{\lambda_3\lambda_4}_{\lambda_1\lambda_2}(p_i)\,, \tag{7.2.16}$$

where $\mathcal{N}$ is a normalization factor, $\lambda_{ij} = \lambda_i - \lambda_j$, and the Wigner d-matrix is

$$d^J_{\lambda\lambda'}(\theta) = \langle \lambda J|\mathrm{e}^{-i\theta J_2}|\lambda' J\rangle\,. \tag{7.2.17}$$

The bras and kets label the elements of the spherical basis of SO(3). Let us study the large J limit of this Wigner d-matrix.

We will understand this limit at the group level. In particular, the raising and lowering operator act on a state as

$$J_3|J\lambda\rangle = \lambda|J\lambda\rangle\,, \qquad J_\pm|J\lambda\rangle = \sqrt{\mathcal{J}^2 - \lambda(\lambda\pm 1)}|J\lambda\pm 1\rangle\,, \tag{7.2.18}$$

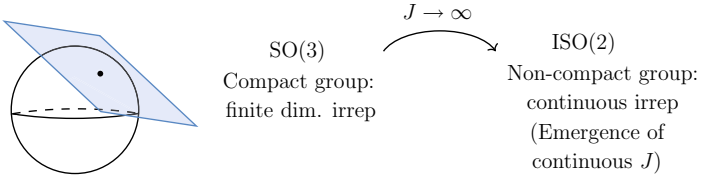
where $\mathcal{J}^2 = J(J+1)$ is the Casimir operator, and the tower contains $2J+1$ states within the irrep. In the limit $\lambda/J \ll 1$, we get

$$J_\pm|J\lambda\rangle = \sqrt{\mathcal{J}^2}|J\lambda\pm 1\rangle + \mathcal{O}(\lambda/J)\,, \tag{7.2.19}$$

where we see that the raising and lowering operators now commute:

$$[J_+, J_-] = 0\,, \tag{7.2.20}$$

and the irreps become infinite dimensional. What we are observing is the contraction of the little group SO(3) to ISO(2), which is a non-compact group and as such admits only infinite dimensional irreps. Hence, at large angular momentum we recover the "flat-earth limit", where the isometries of a sphere reduce to those of a plane:



(7.2.21)

Let us return to the Wigner d-matrix. It would be ideal to find a basis given by the eigenstates of J_+ and J_-, which would drastically simplify the evaluation of the Wigner d-matrix elements (7.2.17). It turns out that such a basis exists and is the continuous-spin basis:

$$j_\pm|\varphi\rangle = \mathrm{e}^{\mp i\varphi}|\varphi\rangle\,, \tag{7.2.22}$$

where φ is an angle. The continuous-spin basis and the $|\lambda\rangle$ basis are connected via a Fourier series, namely

$$|\varphi\rangle \equiv \sum_{\lambda\in(\text{half-})\text{integers}} \mathrm{e}^{i\varphi\lambda}|\lambda\rangle \quad \longleftrightarrow \quad |\lambda\rangle = \int_0^{2\pi} \frac{\mathrm{d}\varphi}{2\pi}\mathrm{e}^{-i\lambda\varphi}|\varphi\rangle\,. \tag{7.2.23}$$

In summary, in the large J limit, the $|J\lambda\rangle$ states can be decomposed into a suitable basis of ISO(2) irreducible representations for which J_2 matrix elements are diagonal. This procedure allows us to recover the Wigner d-matrix $d^J(\theta)$ in the large angular momentum limit, as follows:

$$d^J_{\lambda'\lambda}(\theta) \xrightarrow[J\to\infty]{\theta\to 0} \int_0^{2\pi} \frac{\mathrm{d}\varphi}{2\pi}\mathrm{e}^{i(\lambda'-\lambda)\varphi}\mathrm{e}^{i\theta\mathcal{J}\sin\varphi} = J_{\lambda-\lambda'}(\mathcal{J}\theta)\,, \tag{7.2.24}$$

where $\mathcal{J} \equiv \sqrt{J(J+1)}$ as before. This is just an integral representation of the Bessel $J_\nu(x)$ function.

The Fourier transform emerges naturally from this picture. For example, the integral over the angles θ can be translated into that over q. The Wigner d-matrix becomes the integral over φ as above. In the large J limit, the amplitude exponentiates with the phase shift. More precisely, we have :

$$\underbrace{\mathcal{M}J^{\lambda_3\lambda_4}_{\lambda_1\lambda_2}(s)}_{\frac{\mathrm{e}^{2i\delta_J}-1}{i}} = \mathcal{N}^{-1}\underbrace{\int_{-1}^{1}\mathrm{d}\cos\theta}_{\sim\int_0^\infty \mathrm{d}q\,q}\ \underbrace{d^J_{\lambda_{12}\lambda_{34}}(\theta)}_{\int\frac{\mathrm{d}\varphi}{2\pi}\mathrm{e}^{i(\lambda_{12}-\lambda_{34})\varphi}\mathrm{e}^{i\theta J\sin\varphi}}\ \mathcal{M}^{\lambda_3\lambda_4}_{\lambda_1\lambda_2}(p_i)\,, \tag{7.2.25}$$

This procedure gives a systematic way to extract the phase shift order by order in θ. The regime of the eikonal approximation is

$$\theta = \frac{\alpha_e}{J} \ll 1\,. \tag{7.2.26}$$

Up to $\mathcal{O}(\alpha_e^2)$, it is a simple Fourier transform, but at subleading orders there will be some corrections to this limit of the Wigner d-matrix, which is in principle known to all orders. For example, if we wanted to extract the $\mathcal{O}(\alpha_e^3)$ contribution from the amplitude, we need to slightly modify the 2D Fourier transform to make it consistent with the large J limit of the partial waves.

Notice that the saddle point is dominated by the scaling $J\theta \sim 1$, so at large angular momentum we are automatically in the small angle $\theta \sim t/s \sim \alpha_e/J$ regime. This is the reason why people often directly expand the amplitude at small t/s and refer to this as the "classical limit". In gravity those α_g/J perturbative corrections to log$b$ are precisely the Post-Minkowskian (PM) corrections, as will be illustrated in the next section.

7.2.3 Resumming at $\alpha_e \sim J$ and Bound Orbits

We can now ask what happens when the ratio α_e/J becomes $\mathcal{O}(1)$. It is easy to see that this will correspond to a larger scattering angle, which (spoiler alert) will lead us to bound orbits.

Since in the case of the hydrogen atom we actually know the full solution, we can study that limit explicitly. Recall from the introduction that we are interested in the point-particle limit $b \gg \lambda_{\rm dB} = \frac{1}{|\mathbf{p}|}$, which implies $J = |\mathbf{p}|b \gg 1$. Therefore, this regime must have the following hierarchy: $J \sim \alpha_e \gg 1$ (alternatively, this can be seen by taking the naive $\hbar \to 0$ limit).

We first observe that (up to a J independent phase), the amplitude (7.2.14) exponentiates into the following object

$$S_J \sim \mathrm{e}^{2\frac{i}{\hbar} I(E,J)}\,, \tag{7.2.27}$$

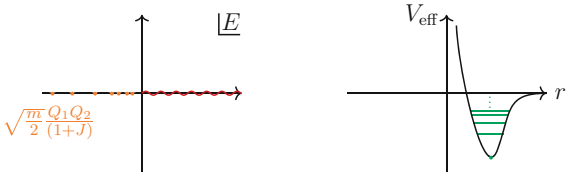
with

$$I(E,J) = \alpha_e \left[-\frac{1}{2}\log(J^2+\alpha_e^2) + 1 - \frac{J}{\alpha_e}\arctan\left(\frac{\alpha_e}{J}\right)\right]. \tag{7.2.28}$$

We wrote down explicitly the $\hbar$ dependence to highlight the fact that it becomes a "classical action". This is a fully classical object usually referred to as the *radial action*.

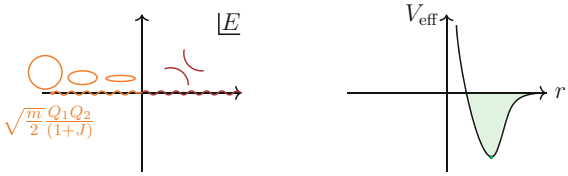
It can be obtained by solving equations of motion for a point particle in an effective potential. Here, we are in the probe limit, so the potential is literally the Coulomb $1/r$ potential generated by the heavy object, but it can be generalized away from the probe limit by constructing an effective potential for one effective body with reduced mass. In Sect. 7.3, we will see more details about this computation in the gravitational case. We will return to the above effective potential and how it can be extracted from the amplitude in a moment.

Before that, let us examine the analytic structure of the amplitude in partial waves $S_J(E)$ from (7.2.14), which looks as follows:



(7.2.29)

It is clear that at negative energy the gamma functions develop some poles, which are the energy levels of the hydrogen atom bound states. In the $\hbar \to 0$ limit, the poles get closer and create a continuum:



(7.2.30)

From the picture of the potential it is easy to see that this should correspond to classical bound orbits. This effect can be seen in the radial action, which develops an imaginary part at negative energies.

Now, let us try to do the analytic continuation explicitly. For this purpose, we write

$$\sqrt{E} = i\sqrt{-E}\,, \tag{7.2.31}$$

with Feynman $i\varepsilon$ prescription, $E \to E + i0^+$ to choose the correct branch. The real part of the action is given by

$$\mathrm{Re}\, I(E, J) = \left[\frac{\pi Q_1 Q_2}{2\sqrt{-2E/m}} + \frac{\pi J}{2}\right]\Theta\Big(-\frac{m Q_1^2 Q_2^2}{2J^2} < E < 0\Big)\,. \tag{7.2.32}$$

This allows us to compute the orbital period and the angle swept after one orbit by taking derivatives with respect to the real part. We find

$$T = 2\frac{\partial\, \mathrm{Re} I}{\partial E} = \frac{\partial \alpha_e}{\partial E} = \frac{\pi Q_1 Q_2 m^{1/2}}{2\sqrt{2}(-E)^{3/2}}\,, \tag{7.2.33}$$

and

$$\Delta\phi = 2\frac{\partial\, \mathrm{Re} I}{\partial J} + \pi = 2\pi. \tag{7.2.34}$$

The extra factor of π is there in order to go from the deflection angle (that we were interested about in the scattering problem) to the angle swept by the particle. Adding subleading corrections can be done using this procedure. We will see later how to derive the perihelion shift in gravity from such considerations.

Audience Question 7.2.1 Why is the time delay supposed to compute the period?
Answer: The statement is that the time delay and the period are analytic continuations of each other. This should not have been obvious a priori. Understanding this type of continuation, in general, remains an open problem.

A fascinating aspect of this derivation is the fact that from a purely classical perspective there is no reason why the bound and unbound regimes are an analytic

continuation of each other or why the choice of ε prescription is the correct one. These features are purely derived from the underlying quantum theory.

7.2.4 Potential, Born Series, and Schrödinger Equation

We obviously do not always have the luxury of easily resumming the amplitudes in momentum space, so how can we access the $\alpha_e \sim J$ regime by computing only a finite number of diagrams in perturbation theory? This section is going to be a bit vague and is meant to give only a general idea behind this problem. This will be made more precise in the following.

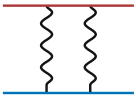
As mentioned earlier, the object we need is the potential V. It is already easy to see that the 3D Fourier transform of the tree-level diagram gives

$$= -m_1 m_2 \int \mathrm{d}^3\mathbf{q}\, \mathrm{e}^{iqr} \frac{Q_1 Q_2}{\mathbf{q}^2}\,, \tag{7.2.35}$$

where $q = |\mathbf{q}|$. Up to a normalization (which we will ignore right now), it is just the Coulomb potential:

$$V(r) = \frac{Q_1 Q_2}{r}\,. \tag{7.2.36}$$

Let us now examine the form of the one-loop integrand in the probe non-relativistic limit. The first step is to solve explicitly for the first component k^0 of the loop momentum $k = (k^0, \mathbf{k})$. Then, expanding the result in the large mass limit, we only get a contribution from the "matter pole" of the form



$$\begin{aligned} &= \int \frac{\mathrm{d}k_0\, \mathrm{d}^3\mathbf{k}}{(2\pi)^4} \frac{1}{(k-p_1)^2 + i\varepsilon} \frac{1}{(k-p_3)^2 + i\varepsilon} \\ &\qquad \times \frac{1}{k^2 - m_1^2 + i\varepsilon} \frac{1}{(p_1+p_2-k)^2 - m_2^2 + i\varepsilon} \\ &\sim \frac{1}{4m_1} \int \frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^3} \frac{Q_1 Q_2}{(p-k)^2} \frac{1}{\frac{\mathbf{k}^2}{2m_2} - E - i\varepsilon} \frac{Q_1 Q_2}{(p+q-k)^2} \\ &\sim \frac{1}{4m_1} \int \frac{\mathrm{d}^3\mathbf{k}}{(2\pi)^3} V((p-k)^2) \frac{1}{\frac{\mathbf{k}^2}{2m_2} - E - i\varepsilon} V((p+q-k)^2)\,, \end{aligned} \tag{7.2.37}$$

which is some sort of iteration of the potential V, confirming that in this limit all the dynamics is hidden in the Coulomb potential. At two loops we get an extra iteration of V, at three loops we get two extra iterations of V, etc.

What is the kernel that controls this iteration? By Fourier transforming the kernel in position space we get

$$G_+(x) \sim \frac{\mathrm{e}^{ip|x-x'|}}{|x-x'|}\,, \tag{7.2.38}$$

which is the retarded Green's function of the wave equation. This should not come as a surprise since in the classical limit the scattering is only influenced by events happening in the past lightcone.

The structure of the amplitude we are observing is nothing but a Born series, which is a perturbative way of solving a wave equation in the small potential regime. We are not going into details here, as this is standard quantum mechanics textbook material (see, for instance, [10]). What is the wave equation whose solution is the non-relativistic hydrogen atom? Obviously, the answer is the Schrödinger equation

$$\left(\frac{\nabla^2}{2m_2} + E\right)\psi = V(r)\,\psi\,, \tag{7.2.39}$$

with $V(r) = Q_1 Q_2/r$.

As a last exercise before moving on to gravity, we can ask how does the radial action emerge from this formulation. Since the radial action is written in the angular momentum space J, we start by rewriting the Schrödinger equation in spherical harmonics:

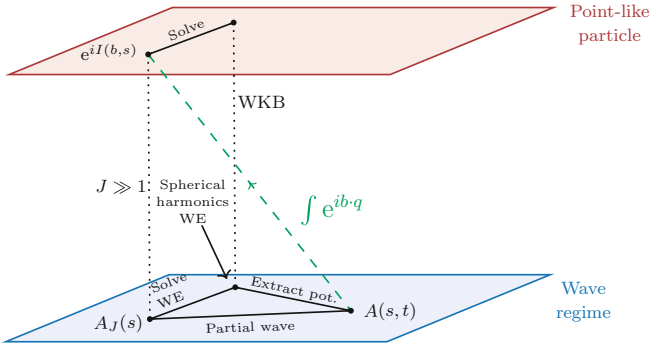
$$-\frac{1}{2m_2 r^2}\frac{\mathrm{d}}{\mathrm{d}r}\left(r^2\frac{\mathrm{d}}{\mathrm{d}r}\psi_J(r)\right) = \left(-E + V(r) + \frac{J(J+1)}{2m_2 r^2}\right)\psi_J(r)\,. \tag{7.2.40}$$

What is the semi-classical limit of the Schrödinger equation? That is precisely the regime of the WKB approximation $\psi_J = e^{\frac{i}{\hbar}I_J(r)}$. Plugging this into the Schrödinger equation, taking the $\hbar \to 0$ limit, and solving for I_J yields

$$I_J(E) = \int \mathrm{d}r \sqrt{2m_2\,(E - V(r)) - \frac{J^2}{r^2}}\,, \tag{7.2.41}$$

which is precisely the radial action that we constructed from the amplitude earlier.

At this stage, let us summarize this section with a graphic representing all the connections we have established:



(7.2.42)

At the level of the wave regime (lower plane on the diagram) we are effectively describing the non-relativistic hydrogen atom (neglecting corrections due to spin, relativistic effects, finite size, etc.). We were able to resum all ladder diagrams into a compact expression (7.2.13). By projecting into partial waves, we obtained (7.2.14), whose poles at negative energies were identified as energy levels of the hydrogen atom. This solution can be obtained directly by solving the Schrödinger equation in spherical harmonics. The dynamics of this problem is fully fixed by the Coulomb potential $V = Q_1 Q_2 / r$ which can be readily extracted from the tree-level amplitude. We saw that in fact ladder diagrams in this regime organise themselves into a Born series of iterations of the Coulomb potential.

From this regime we can take the point particle limit (upper plane of the diagram). At the level of the amplitude this is equivalent to taking the limit $q \ll |\mathbf{p}|$. In this particular example, the form of the amplitude does not change (this is not the case in gravity as we will see next). Similarly, the point particle limit of the amplitude in angular momentum space is reached by taking $J \gg 1$. In this limit the amplitude naturally exponentiates into the radial action. From this object we can extract classical observables (such as scattering angles and time delays) by performing a saddle point analysis. Since in general we do not have access to the full $S_J(E)$ solution, we studied how it can be extracted perturbatively from amplitude calculations, by taking the $J \gg 1$ limit directly at the level of the transform, which up to $\mathcal{O}(\alpha_e^2)$ reduces to a 2D Fourier transform. The radial action can be obtained alternatively from the Schrödinger equation by considering the WKB approximation.

7.3 Gravity and Relativistic Born Series

Miguel Correia

In this section, we are going to generalize what we have seen before to the case of gravity. Recall that we are interested in extracting classical observables from scattering amplitudes. To illustrate how to do this, we are going to review the appearance of some of the classical aspects of GR: bending of light, Shapiro time

delay, precession of the perihelion of Mercury, the effects of horizon dissipation (in Sect. 7.4), emission of gravitational waves (in Sect. 7.5), and non-linear three-body dynamics (in Sect. 7.6). The purpose of this section is to cover the first three effects.

7.3.1 Gravitational Amplitudes

We want to compute scattering amplitudes in the presence of gravity. Let us start with reviewing the basic setup for approaching this computation, as if we lived in the 1950s. The first step is to write down the action, which we can split into:

$$S = S_{\text{gravity}} + S_{\text{matter}} + S_{\text{gauge fixing}} + S_{\text{ghost}} \,. \tag{7.3.1}$$

Let us consider the Einstein–Hilbert action for GR with two minimally coupled scalars Φ_1 and Φ_2:

$$S_{\text{gravity}} = \frac{1}{16\pi G} \int \mathrm{d}^4x \sqrt{-g}\, R \,, \tag{7.3.2}$$

and

$$S_{\text{matter}} = \frac{1}{2} \int \mathrm{d}^4x \sqrt{-g} \left[g^{\mu\nu} \partial_\mu \Phi_1 \partial_\nu \Phi_1 - m_1^2 \Phi_1^2 + g^{\mu\nu} \partial_\mu \Phi_2 \partial_\nu \Phi_2 - m_2^2 \Phi_2^2 \right] . \tag{7.3.3}$$

To construct the Feynman rules, we expand the metric around flat space

$$g_{\mu\nu} = \eta_{\mu\nu} + \sqrt{32\pi G}\, h_{\mu\nu} \,, \tag{7.3.4}$$

and consider $h_{\mu\nu}$ to be the field we want to scatter. This procedure generates higher-point interactions. Schematically, the action looks like

$$S_{\text{gravity}} = \int \mathrm{d}^4x \left[(\partial h)^2 + \sqrt{G} \partial^2 h^3 + \mathcal{O}(G) \right] . \tag{7.3.5}$$

and

$$S_{\text{matter}} = \int \mathrm{d}^4x \left[(\partial \Phi_1)^2 + (\partial \Phi_2)^2 + \sqrt{G} h_{\mu\nu} (T^{\mu\nu}[\Phi_1] + T^{\mu\nu}[\Phi_2]) + \mathcal{O}(G) \right] . \tag{7.3.6}$$

The action has an infinite number of terms, which becomes quite cumbersome at higher orders. This is where the on-shell revolution plays a big role. Essentially, we have learned that we do not need any of this formalism.[2]

[2] At this stage, Miguel crossed out all the above equations.

We are going to follow the "bootstrap" philosophy and use general principles to constrain the amplitudes instead. It will suffice to know that the graviton is a spin-2 massless particle and use a few of the modern tools such as:

- Unitarity and factorization on poles
- Spinor-helicity formalism
- Double copy relations (gravity $\sim$ Yang $-$ $-$Mills^2)
- Generalized unitarity
- Integration by parts and differential equations

In this lecture we are not going to review any of the above methods, instead we are mostly interested in how to extract physical observables from scattering amplitudes.

Analogously to the hydrogen atom case discussed previously, we start with tree-level scattering of $m_1 m_2 \to m_1 m_2$ in gravity, which reads:

$$i\mathcal{M}_0 = \quad = 16\pi G \frac{(2(p_1 \cdot p_2)^2 - m_1^2 m_2^2)}{t} \equiv 16\pi G \frac{m_1^2 m_2^2 (2\sigma^2 - 1)}{t}\,, \tag{7.3.7}$$

in the small t limit. We define $\sigma \equiv \frac{p_1 \cdot p_2}{m_1 m_2}$. The eikonal phase at the leading order in G is then

$$\delta(s, b) = -\alpha_G(s) \log(b/b_{IR})\,, \tag{7.3.8}$$

with

$$\alpha_G(s) = G \frac{2(p_1 \cdot p_2)^2 - m_1^2 m_2^2}{\sqrt{s}|\mathbf{p}|} = \frac{G m_1 m_2 (2\sigma^2 - 1)}{\sqrt{\sigma^2 - 1}}\,. \tag{7.3.9}$$

In the non-relativistic limit (say one of the particles is very heavy, $m_1 = M \gg m = m_2$) we have $p_1 \cdot p_2 = Mm$ and $\sqrt{s} = M$. This gives us

$$\alpha_G = \frac{2GMm^2}{|\mathbf{p}|} = \frac{2GMm}{\sqrt{2E/m}}\,, \tag{7.3.10}$$

where $E \equiv \frac{|\mathbf{p}|^2}{2m}$ coming from $E \equiv \sqrt{s} - M - m$. Notice that structurally, this is the same effective coupling we got before in (7.2.9) for QED.

Let us now consider a classical test of GR: bending of light. When one of the particles is massless, we have $\sqrt{s} = M + \omega$ and $p_1 \cdot p_2 = M\omega$. This yields

$$\alpha_G = 2GM\omega\,. \tag{7.3.11}$$

The deflection angle is given by $\theta = 2\frac{\partial\delta(s,b)}{\partial J}$, where $b = J/\omega$. This gives us

$$\theta = \frac{4GM\omega}{J} = \frac{4GM}{b}\,, \tag{7.3.12}$$

which is the classic deflection angle formula in GR. Notice that the deflection angle does not depend on the frequency ω, just on the impact parameter b.

Audience Question 7.3.1 Is the fact that ω cancels out a consequence of the equivalence principle?
Answer: Yes, exactly. Higher orders including quantum corrections will give dependence on the frequency.

We can also consider the Shapiro time delay. Recall that the time delay is given by $\Delta t = 2\frac{\partial\delta(s,b)}{\partial\sqrt{s}}$. In the case of a null geodesic, for which $\sqrt{s} = \omega + M$, we get

$$\Delta t = -4GM\log(b/b_{IR}) > 0\,, \tag{7.3.13}$$

which again is the correct answer in GR. As expected, gravity slows you down.

Let us now do something slightly more interesting and ask what would happen as we go to higher orders in the expansion. The amplitude admits an expansion with more and more loops:

$i\mathcal{M} =$  $+\ldots$ (7.3.14)

In the hydrogen atom, the tree-level diagram contains the Coulomb potential and the ladder diagrams are iterations of it. However, in GR we have mixing between different terms. To distinguish which terms contribute to the 'potential' and which terms are iterations, we will use the connection to an "effective-one-body" Lippmann–Schwinger equation, which resums all such contributions. We will follow the exposition of Todorov [3], who did his work in 1970 while at IAS (see also [9] for a modern review).

7.3.2 Relativistic Born Series

In the center of mass frame we have the ingoing momenta

$$p_1^\mu = (E_1, \mathbf{p})\,, \qquad E_1 = \sqrt{|\mathbf{p}|^2 + m_1^2}\,, \tag{7.3.15}$$

$$p_2^\mu = (E_2, -\mathbf{p})\,, \qquad E_2 = \sqrt{|\mathbf{p}|^2 + m_2^2}\,, \tag{7.3.16}$$

with the outgoing momenta given by

$$p_3^\mu = (E_1, \mathbf{p}') = (E_1, \mathbf{p} + \mathbf{q}) \,, \tag{7.3.17}$$

$$p_4^\mu = (E_2, -\mathbf{p}') = (E_2, -\mathbf{p} - \mathbf{q}) \,. \tag{7.3.18}$$

Here, $\mathbf{q} = \mathbf{p}' - \mathbf{p}$ is the momentum exchange. Let $T(s,t)$ be the scattering amplitude for the scattering of two scalars of mass m_1 and m_2 where the Mandelstam invariants read

$$s = (p_1 + p_2)^2, \qquad t = (p_1 - p_3)^2 \,. \tag{7.3.19}$$

These relate to $\mathbf{p}$ and $\mathbf{q}$ in the center of mass frame via

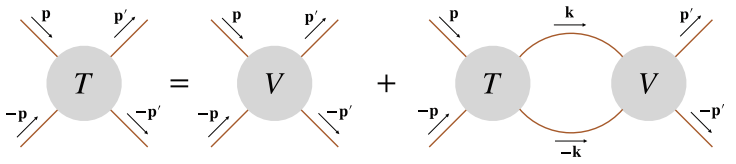
$$|\mathbf{p}|^2 = \frac{\left[s - (m_1 + m_2)^2\right]\left[s - (m_1 - m_2)^2\right]}{4s}, \qquad t = -|\mathbf{p}' - \mathbf{p}|^2 = -|\mathbf{q}|^2. \tag{7.3.20}$$

The vectors $\mathbf{p}$ and $\mathbf{q}$ are further constrained by the condition $\mathbf{p} \cdot \mathbf{q} = -|\mathbf{q}|^2/2$ that stems from energy conservation $|\mathbf{p}| = |\mathbf{p}'|$. We will use the notation $T(\mathbf{p}, \mathbf{p}')$ where it is understood that the dependence on s and t is given in terms of $\mathbf{p}$ and $\mathbf{p}'$ in Eq. (7.3.20).

We define the *potential* $V(\mathbf{p}, \mathbf{q})$ in terms of the Lippmann–Schwinger equation:

$$T(\mathbf{p}, \mathbf{p}') = V(\mathbf{p}, \mathbf{p}') + \int \mathrm{d}^3\mathbf{k}\, T(\mathbf{p}, \mathbf{k}) G(\mathbf{p}, \mathbf{k}) V(\mathbf{k}, \mathbf{p}') \,, \tag{7.3.21}$$

where $G(\mathbf{p}, \mathbf{k})$ is a Green's function. It is not unique, but it will have to satisfy some consistency conditions and we will fix it shortly. This equation is easier to understand at the level of pictures:



(7.3.22)

Note that in the last term, the momentum $\mathbf{k}$ is off-shell and we integrate over it.

The fact we chose this to be a 3D integral makes it connect directly to the standard Lippmann–Schwinger equation of one-body quantum mechanics. In fact, in the eikonal or WKB limit one finds a relation to the effective one-body formalism by Buonanno and Damour [11, 12].[3]

[3] Miguel was not sure about the spelling of Alessandra Buonanno's family name. The Italians in the audience advised that it should contain three instances of the letter n.

We now require the potential $V(\mathbf{p}, \mathbf{p}') = V(s,t)$ to be a real function in the scattering regime $s \geqslant (m_1 + m_2)^2$, i.e.,

$$\mathrm{Im} V(\mathbf{p}, \mathbf{p}') = 0 \qquad \text{for} \qquad s \geqslant (m_1 + m_2)^2 \,. \tag{7.3.23}$$

We will now also require consistency of the Lippmann–Schwinger equation with elastic unitarity in order to fix the form of $G(\mathbf{p}, \mathbf{k})$. Elastic unitarity, which is the optical theorem applied below the three-particle threshold, reads

$$\begin{aligned} \mathrm{Im}\, T(\mathbf{p}, \mathbf{p}') &= \frac{1}{8\pi^2} \int \mathrm{d}^4 k \; \delta^+\big(k^2 - m_1^2\big)\delta^+\big((k - p_1 - p_2)^2 - m_2^2\big) T(p,k)\, T^*(k, p') \\ &= \frac{1}{16\pi^2 \sqrt{s}} \int \mathrm{d}^3 \mathbf{k}\, \delta(|\mathbf{k}|^2 - |\mathbf{p}|^2)\, T(\mathbf{p}, \mathbf{k})\, T^*(\mathbf{k}, \mathbf{p}') \,, \end{aligned} \tag{7.3.24}$$

where k is the 4-momentum and δ^+ denote putting the corresponding particle on-shell and imposing the positive-energy condition. In the second line, we recast it as a 3D integral in order to resemble the Lippmann–Schwinger equation, by getting rid of one of the delta functions. Since we imposed that $V(\mathbf{p}, \mathbf{q})$ is real across the 2-particle cut, the discontinuity of the amplitude should be captured by the Green's function $G(\mathbf{p}, \mathbf{k})$.

To arrive at the constraint on G as quickly as possible, it will pay off to work with a short-hand notation. Schematically, the relativistic Born series can be written as

$$T = V + VGV + VGVGV + \cdots = V \frac{1}{1 - GV} \,, \tag{7.3.25}$$

where multiplication denotes integrating over the intermediate momentum, as above, and we suppress all the constants and kinematic dependence. Imposing the reality condition, $V^* = V$, gives

$$T^* = V \frac{1}{1 - G^* V} \,. \tag{7.3.26}$$

In the same notation, Elastic unitarity takes the form

$$T - T^* = T T^* \,. \tag{7.3.27}$$

Using (7.3.25) and (7.3.26), the right-hand side can be written as

$$T T^* = V \frac{1}{1 - GV} V \frac{1}{1 - G^* V} \,, \tag{7.3.28}$$

whereas the left-hand side is

$$T - T^* = V\left[\frac{1}{1-GV} - \frac{1}{1-G^*V}\right] = V\frac{1}{1-GV}(G-G^*)V\frac{1}{1-G^*V}\,. \tag{7.3.29}$$

Unitarity then fixes $G - G^* = 1$, i.e., the imaginary part of the Green's function is a delta function. After carefully working out all the factors, in terms of equations this constraint reads

$$\mathrm{Im}\, G(\mathbf{p},\mathbf{k}) = \frac{\delta(|\mathbf{k}|^2 - |\mathbf{p}|^2)}{16\pi^2\sqrt{s}}\,. \tag{7.3.30}$$

The simplest choice is to write the solution as

$$G(\mathbf{p},\mathbf{k}) = \frac{1}{(2\pi)^3}\frac{1}{2\sqrt{s}}\frac{1}{|\mathbf{k}|^2 - |\mathbf{p}|^2 - i\epsilon}\,, \tag{7.3.31}$$

This solution is written up to analytic (real) terms which can be absorbed into the potential.

Audience Question 7.3.2 So what exactly is the freedom in choosing G?
Answer: You can add any analytic function to the above solution, as long as it is compatible with the constraint (7.3.30). Different choices will lead to other effective potentials V differing by off-shell pieces. In coordinate space these choices correspond to different coordinate systems [9]. We observe that the above choice selects isotropic coordinates (see below).

Let us take the non-relativistic probe limit as a cross-check. This amounts to setting:

$$\sqrt{s} = M + m + E\,, \qquad |\mathbf{p}|^2 = 2\mu E = \frac{2mM}{m+M}E\,, \tag{7.3.32}$$

where we made use of Eq. (7.3.20) and $\mu = \frac{m_1 m_2}{m_1+m_2}$ is the reduced mass of the system. In this limit, we get

$$G \sim \frac{1}{\frac{|\mathbf{k}|^2}{2\mu} - E - i\varepsilon}\,. \tag{7.3.33}$$

This is precisely the non-relativistic propagator from the Schrödinger equation, as expected.

Let us examine this Green's function closer. The $|\mathbf{k}|^2$ dependence is the same in the relativistic and non-relativistic cases. It means that after a Fourier transform, we still get quadratic derivatives in the spatial directions. The difference is the dependence on E, which is linear in the non-relativistic case. This is in contrast to the relativistic case, in which we had

$$|\mathbf{p}|^2 = \frac{[s-(m_1+m_2)^2][s-(m_1-m_2)^2]}{4s}\,, \tag{7.3.34}$$

and the energy dependence is very non-linear. To summarize, we end up with an effective Schrödinger equation with

$$|\mathbf{k}|^2 \to -\nabla^2\,, \qquad |\mathbf{p}|^2 = f(E)\,, \tag{7.3.35}$$

for $f(E)$ given by Eq. (7.3.34) where $s = E^2$. This is called the relativistic effective one-body Schrödinger equation:

$$f(E)\Psi = (\nabla^2 + V)\Psi\,. \tag{7.3.36}$$

Similar manipulations can of course also be done for the hydrogen atom. By drawing more and more complicated diagrams, one can compute things like the QED corrections to the spectrum, vacuum polarization, anomalous magnetic moment of the electron, positronium lifetime, etc. [3].

The summary so far is as follows. We first wrote down the Lippmann–Schwinger equation. Then we found that the only way it can be consistent with elastic unitarity with V being real is that G has the form given above. After this is done, we can use a perturbative expansion in terms of Feynman diagrams to fix the potential V and compute classical observables.

7.3.3 Potential at $\mathcal{O}(G^2)$, Perihelion Precession, and Higher Orders

We may now go ahead and compute loop corrections to the potential by matching the Feynman diagram expansion with the Born series of the Lippmann–Schwinger equation in Eq. (7.3.21).

For example, at one-loop we have:

(7.3.37)

In this case, the cross-box (second diagram) contributes at leading order in the classical, (or small q), limit and cancels part of the box (first diagram). This is

a crucial difference between defining the potential via the Lippmann–Schwinger equation (7.3.21). In the latter case only the box contributes as an "iteration" while the crossed-box gets absorbed in the potential, which makes the potential ill-defined.[4]

Performing the full computations at one-loop order gives use the $\mathcal{O}(G^2)$ terms:

$$\mathcal{M}_2 = \frac{6\pi^2 G^2 (m_1+m_2)(5(p_1\cdot p_2)^2 - m_1^2 m_2^2)}{|\mathbf{q}|} + \int \mathrm{d}^3\mathbf{k}\, \mathcal{M}_1(\mathbf{p},\mathbf{k}) G(\mathbf{p},\mathbf{k}) \mathcal{M}_1(\mathbf{k},\mathbf{p}') \tag{7.3.38}$$

where

$$\mathcal{M}_1(\mathbf{p},\mathbf{k}) = -\frac{8\pi G\left(m_1^2 m_2^2 - 2(p_1\cdot p_2)^2\right)}{|\mathbf{q}|^2}\,, \tag{7.3.39}$$

is the tree-level amplitude.

Matching it to the Born series leads to the effective potential

$$\begin{aligned} V(\mathbf{p},\mathbf{q}) = &-\frac{8\pi G\left(m_1^2 m_2^2 - 2(p_1\cdot p_2)^2\right)}{|\mathbf{q}|^2} \\ &- \frac{6\pi^2 G^2 (m_1+m_2)(m_1^2 m_2^2 - 5(p_1\cdot p_2)^2)}{|\mathbf{q}|} + \mathcal{O}(G^3)\,. \end{aligned} \tag{7.3.40}$$

The Fourier transform to position space gives

$$V(r) = \frac{G}{r}\frac{m_1^2 m_2^2 - 2(p_1\cdot p_2)^2}{\sqrt{s}} - \frac{G^2}{r^2}\frac{3m_1^2 m_2^2 (m_1+m_2)}{2\sqrt{s}}(1-5\sigma^2) + \mathcal{O}(G^3)\,, \tag{7.3.41}$$

where we remind that

$$\sigma \equiv \frac{p_1\cdot p_2}{m_1 m_2}\,. \tag{7.3.42}$$

We can then compute $\mathcal{O}(G^2)$ corrections to the deflection angle. In fact, the state of the art is currently $\mathcal{O}(G^4)$, or equivalently 3 loops. For the purposes of this section, we will be satisfied with $\mathcal{O}(G^2)$.

[4] In his work [3], Todorov writes: "I wish to thank Professor F. Dyson for an enlightening discussion prior to this work, and especially for his refusal to be satisfied with any two-particle equation which does not lead [in the probe limit $m_1 \ll m_2$] to the Klein–Gordon (or Dirac) equation in an external field."

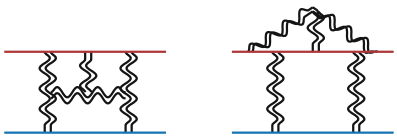
Let us compute one more classical test of GR: the precession of the perihelion of Mercury. It is reproduced by plugging the above potential into the radial action (7.2.41) and taking its J derivative:

$$\Delta\Phi = \pi + \frac{\partial I_J(E)}{\partial J} = G^2 \frac{3\pi m_1^2 m_2^2 (m_1+m_2)}{2J^2\sqrt{s}}(5\sigma^2 - 1)\,. \tag{7.3.43}$$

In the non-relativistic limit we have $\sigma \to 1$ and $\sqrt{s} \to m_1 + m_2$. If in addition, we use the probe limit $m = m_2 \ll m_1 = M$ and make use of the identity $J^2 = m^2 GM(1-e^2)a$, we find the famous result of the Mercury perihelion shift

$$\Delta\Phi_{\text{Mercury}} = \frac{6\pi GM}{(1-e^2)a}\,. \tag{7.3.44}$$

Now, let us say some words about higher orders in this perturbative expansion in G. What is exciting about this expansion is that there is some new physical effect at every order. At $\mathcal{O}(G^2)$ we have the perihelion precession just discussed. At order $\mathcal{O}(G^3)$ there are radiation effects. For radiative effects one can use the so-called Kosower–Maybee–O'Connell (KMOC) formalism [13]. The big open problem in this area is how to analytically continue the result of the scattering problem to the bound-state problem. The problem occurs at $\mathcal{O}(G^4)$, where you get diagrams of this form:



(7.3.45)

These are responsible for the *tail effects*. The radiated graviton feels the attraction of the two-body system it was emitted from (this does not happen in QED due to lack of photon self-interactions). When trying to compute the effective potential with these diagrams, we get terms of the form $G^4\rho(E)$, where $\rho(E)$ is a distribution, not a function, of the energy. It does not admit a good analytic continuation. The way to deal with it is a big open problem.

7.3.4 Wave Regime and Regge–Wheeler Equation

Let us now consider the wave regime mentioned in Sect. 7.2, as presented in the bottom right corner of Eq. (7.1.2), which is the relevant regime for the ringdown. In this case, we have one massless particle with frequency ω scattering off a massive particle with mass M. Hence, the energy is $\sqrt{s} = M + \omega$ with $\omega \ll M$. We are in the regime in which the impact parameter $b \sim \lambda \sim \frac{1}{\omega} \gg \lambda_C$, where λ_C is the Compton wavelength.

In this case, we can compute the potential once again, but this time the triangle diagram gives an extra contribution. After the dust settles, we get:

$$V(\mathbf{p},\mathbf{q}) = \frac{16\pi G M^2\omega^2}{|\mathbf{q}|^2} + \frac{15\pi^2 G^2 M^3\omega^2}{|\mathbf{q}|} + \frac{G^2M^2\pi^2}{2}\frac{\mathbf{p}\cdot\mathbf{q}}{|\mathbf{q}|} + \dots \tag{7.3.46}$$

The first term is the tree-level contribution. The second term is precisely what comes from the triangle diagram. The third and later terms are subdominant in the point-particle limit.

What does this potential give after the Fourier transform? The answer is

$$V(r) = \left(\frac{4MG\omega^2}{r} + \frac{15G^2M^2\omega^2}{2r^2}\right) + \frac{G^2M^2}{2r^3}\hbar^2\partial_r + \dots \tag{7.3.47}$$

Only the first two terms are relevant. The corresponding wave operator is

$$|\mathbf{p}|^2 - |\mathbf{k}|^2 \to \omega^2 - \nabla^2\,, \tag{7.3.48}$$

and the associated Schrödinger equation reads

$$(\omega^2 - \nabla^2)\phi = V\phi\,. \tag{7.3.49}$$

What is this equation? It turns out to be the Regge–Wheeler equation in isotropic coordinates. Let us confirm this.

We are going to start with the action for a scalar field on a gravitational background, which reads

$$S = \int \mathrm{d}^4x\sqrt{-g}\, g^{\mu\nu}\partial_\mu\phi\,\partial_\nu\phi\,. \tag{7.3.50}$$

The wave equation follows from

$$\frac{\delta S}{\delta\phi} = 0 \qquad \Longrightarrow \qquad \partial_\mu(\sqrt{-g}\, g^{\mu\nu}\partial_\nu\phi) = 0. \tag{7.3.51}$$

The Schwarzschild metric, in isotropic coordinates, is written as

$$g_{\mu\nu} = \mathrm{diag}\big(A(r), -B(r), -r^2B(r), -r^2\sin^2\theta B(r)\big)\,, \tag{7.3.52}$$

or equivalently

$$\mathrm{d}s^2 = A(r)\mathrm{d}t^2 - B(r)\big[\mathrm{d}r^2 + r^2\mathrm{d}\theta^2 + r^2\sin^2\theta\,\mathrm{d}\varphi^2\big]\,, \tag{7.3.53}$$

with

$$A(r) = \frac{\left(1 - \frac{GM}{2r}\right)^2}{\left(1 + \frac{GM}{2r}\right)^2}, \qquad B(r) = \left(1 + \frac{GM}{2r}\right)^4. \tag{7.3.54}$$

Noting that $\sqrt{-g} = \sqrt{AB^3}\, r^2 \sin\theta$, the wave equation on this background is given by

$$\frac{1}{A}\partial_t^2\phi - \frac{1}{B}\nabla^2\phi - \frac{1}{\sqrt{AB^3}}\partial_r\left(\sqrt{AB}\right)\partial_r\phi = 0\,, \tag{7.3.55}$$

where $\nabla^2 = \eta^{ij}\partial_i\partial_j$ is the Euclidean Laplacian. Rearranging terms we obtain

$$\nabla^2\phi = \left(\frac{B}{A}\partial_t^2 - \frac{\partial_r(AB)}{2AB}\partial_r\right)\phi\,. \tag{7.3.56}$$

Next, we write it in terms of the operators $\hat{E} = -i\hbar\partial_t$ and $\hat{\mathbf{p}} = -i\hbar\boldsymbol{\partial}$, and replace A and B from (7.3.54). Lastly, by matching to the wave equation (7.3.49), we can extract the potential at all orders in G

$$V = \left(\frac{\left(1 + \frac{GM}{2r}\right)^6}{\left(1 - \frac{GM}{2r}\right)^2} - 1\right)\partial_t^2 + \frac{2G^2M^2}{4r^3 - G^2M^2r}\hbar^2\partial_r\,. \tag{7.3.57}$$

Expanding the potential up to $\mathcal{O}(G^2)$ we obtain

$$V = \left(\frac{4MG}{r} + \frac{15G^2M^2}{2r^2}\right)\omega^2 + \frac{G^2M^2}{2r^3}\hbar^2\partial_r + \mathcal{O}(G^3)\,. \tag{7.3.58}$$

The Fourier Transform of this expression precisely matches (7.3.46).

Notice that the first bracket dominates at high frequencies, which corresponds to the Eikonal regime described earlier. We can indeed recognise the corrections up to $\mathcal{O}(G^2)$ to the light bending problem that enter in the relevant radial action and allow one to extract the classical scattering angle $\theta = \frac{2GM}{b} + \frac{15\pi G^2M^2}{8b^2}$.

We may use this understanding to compute gravitational amplitudes for massless particles scattering off black hole backgrounds via the Born series and Fourier transforms, instead of standard Feynman integrals. We envision two possible applications:

- **Scattering off a Kerr black hole** Current interest in the community is to compute the tree-level $\mathcal{O}(G)$ scattering amplitude off the Kerr black hole in the classical wave regime. We believe that this Kerr–Compton amplitude is simply the Born amplitude, which can be computed by a Fourier transform of the Kerr potential [9].

- **Black hole tidal Love numbers** The next section will describe how to account for dissipative and tidal effects in terms of Feynman diagrams. We envision that the matching calculation between effective field theory (EFT) and the exact solution from GR can be done at the level of the Born series more efficiently than via the use of Feynman diagrams. We expect that tidal Love numbers are encoded as 'delta function' contributions in the potential (https://inspirehep.net/literature/2901479).

7.4 Worldline EFT, Astrophysics Applications, and Gravitational Raman Scattering

Zihan Zhou

7.4.1 Worldline EFT and Astrophysics Applications

In the previous sections, we mostly calculated the scattering amplitude of binary dynamics within the point particle approximation. While this is a good approximation when the two compact objects are far away, it often fails to accurately model the internal structures such as the tidal response of neutron stars and black holes. To address these limitations and incorporate more complex physical effects, we turn to the worldline EFT, which utilizes a multipole expansion approach analogous to that used in electromagnetism [14]. The multipole expansion in worldline EFT is structured as a series where higher order terms capture more detailed angular information about the system. The simplest term is the monopole, represented by m. To account for the object's structure more comprehensively, we introduce higher-order terms: dipole P_i, quadrupole Q_{ij}, octupole Q_{ijk}, and potentially higher multipoles. These terms allow us to model the detailed structure of compact objects beyond the basic point particle description.

Based on this decomposition, we can write down the effective action within a given gravitational background [15–18]

$$S_{\rm pp} = -\int \mathrm{d}\tau \left[m + Q_L^E E^L + \dots \right] + (E \to B)\,, \tag{7.4.1}$$

where $L = i_1 i_2 \cdots i_\ell$ shows the multipole indices. Similar to what we have learned in electromagnetism, in gravitational contexts E^L and B^L represent the electric and magnetic components of the tidal fields, respectively. Each component exhibits distinct parity transformation properties. These fields are defined as follows:

$$E_L = \begin{cases} \partial_{\langle L\rangle}\phi & \text{if} \quad s=0\,, \\ \partial_{\langle L-1\rangle}E_i & \text{if} \quad s=1\,, \\ \partial_{\langle L-2\rangle}E_{ij} & \text{if} \quad s=2\,. \end{cases} \qquad B_L = \begin{cases} \partial_{\langle L-1\rangle}B_i & \text{if} \quad s=1\,, \\ \partial_{\langle L-2\rangle}B_{ij} & \text{if} \quad s=2\,, \end{cases} \tag{7.4.2}$$

where s denotes the spin of the external field. Let us use the notation $\langle \cdots \rangle$ to represent the symmetric-trace-free component of tensors. In our discussion on gravitational interactions (spin-2 fields), we consider a point particle with its four-velocity, u^μ, and tetrads e_i^μ. The definitions of the electric and magnetic tidal fields are given by

$$E_{ij} = u^\mu e_i^\nu u^\rho e_j^\sigma C_{\mu\nu\rho\sigma} \ , \quad B_{ij} = u^\mu e_i^\nu u^\rho e_j^\sigma \star C_{\mu\nu\rho\sigma} \ , \tag{7.4.3}$$

where $C_{\mu\nu\rho\sigma}$ is the Weyl tensor and $\star C_{\mu\nu\rho\sigma}$ stands for its dual. One can readily verify that the effective action presented in Eq. (7.4.1) is both gauge-invariant and re-parametrization invariant.

From the EFT perspective, we are going to treat $Q_L^{E/B}$ as composite operators that depend on some microscopic variables X

$$S_{\rm pp} \to S_{\rm pp} + \int {\rm d}\tau \mathcal{L}_Q[Q_L^{E/B}(X), \dot{Q}_L^{E/B}(X)] \, . \tag{7.4.4}$$

For a fluid star, the microscopic variable X corresponds to the Lagrangian displacement ξ of the fluid element, $X = \xi$ [19]. With this basic knowledge in mind, let us try to apply this theoretical framework in some astrophysical context.

Audience Question 7.4.1 How much of an approximation is treating stars as fluid stars? For example, is the Sun a fluid star?
Answer: It is typically a very good approximation. Our Sun can be treated as a fluid star with polytropic index $n = 3$.

7.4.1.1 Application 1: Stellar Oscillation, Tidal Encounters and Tidal Disruption Events

The first application is to use worldline EFT to model stellar oscillations, specifically to analyze tidal encounters and tidal disruption events. For simplicity, we will focus on the Newtonian limit and assume a linear tides approximation. Part of the discussion in this section comes from ongoing work with J. Li, I. Martínez-Rodríguez.

By varying the action in Eq. (7.4.4) with respect to the composite variable Q_L^E, we can get the following Euler-Lagrangian equation that governs the evolution of the multipole moments

$$\frac{\rm d}{{\rm d}\tau}\Big(\frac{\partial \mathcal{L}_Q}{\partial \dot{Q}_L^E}\Big) - \frac{\partial \mathcal{L}_Q}{\partial Q_L^E} = -E^L \, . \tag{7.4.5}$$

Although the equation presented above may seem quite abstract at this stage, we can still employ the method of retarded Green's functions

$$\langle Q_L^E(\tau) Q_{L'}^E(\tau')\rangle_{\rm ret} = i\langle [Q_L^E(\tau), Q_{L'}^E(\tau')]\rangle \Theta(\tau-\tau') \equiv \delta_{LL'} F_\ell(\tau-\tau') \, , \tag{7.4.6}$$

to solve for the evolution of Q_L^E

$$\langle Q_L^E(\tau)\rangle = -\int_{-\infty}^{\tau} \mathrm{d}\tau' F_\ell(\tau-\tau') E_L(\tau') \,. \tag{7.4.7}$$

To better understand the structure of this retarded Green's function, we will use the spectral representation, which decomposes the full response function into a summation of eigenmodes ω_n accompanied by their corresponding overlap integrals $\langle n|Q_L|0\rangle$

$$\mathrm{Re}F_\ell(\omega) = -\sum_n |\langle n|Q_L|0\rangle|^2 \mathrm{PV}\frac{2\omega_n}{\omega^2-\omega_n^2} \,, \tag{7.4.8}$$

where PV here stands for the principle value. Now, once we have determined the eigenmode frequencies and the overlapping integral, we can obtain the evolution of the quadrupole moments. However, the calculation of eigenmodes falls outside the scope of the EFT; instead, these must be obtained through matching the EFT with an UV theory. For our purposes, this UV theory is the stellar perturbation theory. In Newtonian gravity, linear fluid perturbations are well-understood, and the corresponding response function is given by Eq. (3.29) in [20].

Now, turning to astrophysical applications, let us examine the scenario of a star and black-hole encounter. Consider a star on a parabolic orbit around a supermassive black hole, with a pericenter distance r_p much greater than the Schwarzschild radius r_s, allowing us to disregard relativistic effects. We want to ask the question: how much energy does the star gain during this process? Based on our intuition from the driven harmonic oscillators, the answer is just the work done by the external tidal field

$$\langle \mathcal{E}_Q\rangle(\tau) = -\int_{-\infty}^{\tau} \mathrm{d}\tau' \langle \dot{Q}_L^E(\tau')\rangle E^L(\tau') \,. \tag{7.4.9}$$

We call $\mathcal{E}_Q$ the tidal energy resulting from the star and black-hole encounter. In most cases, this tidal energy is less than the star's binding energy, leading to the star experiencing only minor oscillations after the encounter. However, there are exceptions where, for some unlucky stars, the tidal energy during the encounter greatly exceeds their binding energy U, i.e.

$$\mathcal{E}_Q \gg |U| \,. \tag{7.4.10}$$

In this case, the star cannot be stable anymore, and will break apart in the end. This is known as the tidal disruption event (TDE). Currently, we have observed around 100 such events [21–23].

7.4.1.2 Application 2: Gravitational Wave Data Analysis

The second application is to put constraints on the tidal deformation and dissipation coefficients from the current LIGO–Virgo–KAGRA data. The candidates in the current data catalog are mostly quasi-circular binary systems. During the inspiral phase, the wavelength of the gravitational waves is significantly larger than the radius of the black hole. This allows us to apply a low-frequency expansion to the retarded response function. Specifically, we will concentrate on the non-spinning quadrupole sector $\ell = 2$:

$$\langle Q_{ij} Q_{kl} \rangle_{\rm ret}(\omega) = -M(GM)^4 \left[\Lambda + i(GM\omega) H_\omega + (iGM\omega)^2 \Lambda_{\omega^2} + \dots \right] \delta_{\langle ij \rangle, \langle kl \rangle} \,. \tag{7.4.11}$$

In the above expansion, Λ is the static Love number $\sim (\frac{R}{GM})^5$ ($\Lambda \sim \left(\frac{R}{GM}\right)^{2\ell+1}$ for general multipole sectors). H_ω is known as the dissipation number because it relates to the time-reversal odd component of the response function. In astrophysical literature, this is often referred to as tidal heating, as the orbital binding energy is converted into heat within the star. Λ_{ω^2} is the dynamical Love number, reflecting the frequency-dependent tidal response. While the impact of static Love numbers on waveforms has been extensively studied for both black holes and neutron stars, the aspect of dissipation has not received much attention. Here, we will briefly summarize our latest results from [24].

At the level of gravitational waveforms, introducing dissipation leads to dephasing effects. Specifically, as shown in Fig. 7.2, we provide a direct comparison between the standard IMRPhenomD waveform (which describes

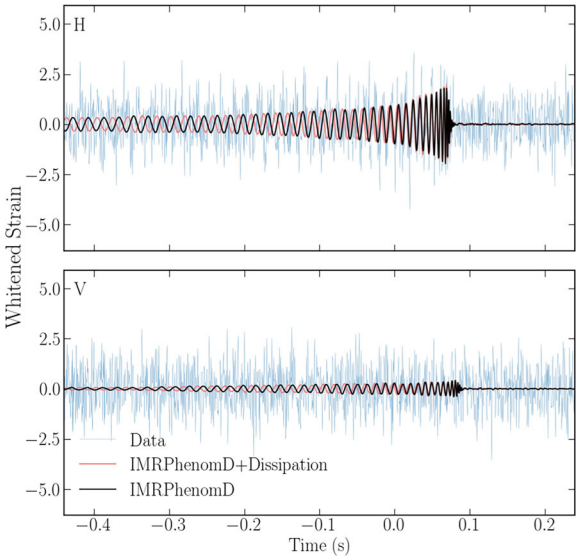


Fig. 7.2 GW strains of the IMRPhenomD and IMRPhenomD+Dissipation waveforms for a GW191216_213338 event in the Hanford and Virgo data. We choose the individual dissipation parameters $H_{1\omega} = H_{2\omega} = 10$ to clearly illustrate dephasing of the waveform due to tidal heating. Reprinted with permission from [24]. © 2024, The Author(s). All rights reserved

the inspiral-merger-ringdown waveform for aligned-spin, quasi-circular binary black holes) and the IMRPhenomD waveform modified to include dissipation (IMRPhenomD+dissipation). While these two waveforms are synchronized at the merger point, they begin to deviate from each other as we trace them back to the earlier inspiral phase. After conducting a detailed Bayesian analysis, we have established constraints on the mass-weighted dissipation number $\mathcal{H}_0 \equiv \frac{1}{M^4}\left(m_1^4 H_{1\omega} + m_2^4 H_{2\omega}\right)$ which range from $-13 < \mathcal{H}_0 < 20$ at 90% confidence level (CL). Additionally, we have constrained the ratio of the energy lost due to tidal dissipation $\Delta E_{\rm H}$, to the radiative energy observed at infinity ΔE_∞, finding that $-0.0026 < \Delta E_{\rm H}/\Delta E_\infty < 0.0025$ at 90% CL. Our results are consistent with GR predictions, which suggest that $\mathcal{H}_0 = 2/45$ for equal mass binary black holes and $\Delta E_{\rm H}/\Delta E_\infty \simeq (10^{-5}, 10^{-4})$.

7.4.2 Gravitational Raman Scattering

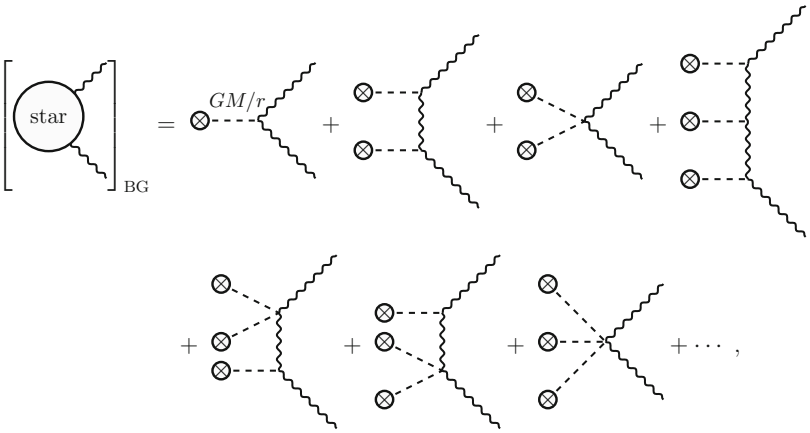
As discussed, worldline EFT proves highly valuable in studying tidal effects within astrophysical contexts. Tidal Love numbers and dissipation numbers can, in principle, be constrained using observational data. However, the analysis becomes significantly more complicated for relativistic compact objects, as it involves systematic consideration of relativistic corrections. To achieve a high precision study of tidal effects and relativistic corrections, let us consider the following massless wave scattering on the stellar background, which we call gravitational Raman scattering [25]:

(7.4.12)

This type of scattering amplitude is particularly interesting because it allows for a perturbative computation from the perspective of EFT, while from the UV side, the amplitude is encoded in the solutions to the linear perturbations of stars. Specifically, for 4D black holes, the perturbation equation, known as the Teukolsky equation [26], can be solved quasi-exactly [27–30]. Solving this equation provides a way to uncover various properties of black holes.

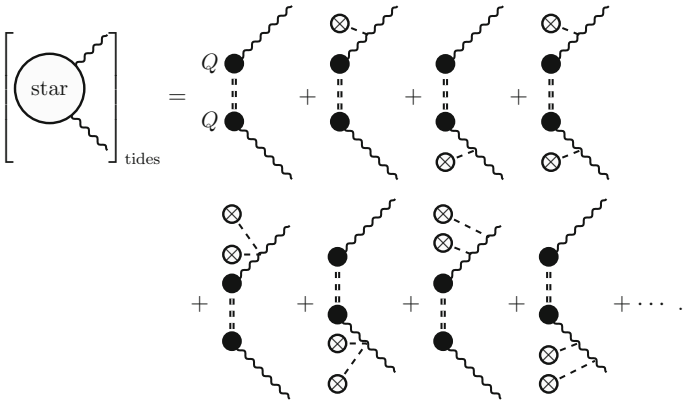
Essentially, there are two regimes in which one can study this problem. In the eikonal limit where $GM\omega \gg 1, \omega b \gg 1$ as discussed in Sect. 7.3, the analysis approximates null geodesics. However, this regime cannot probe the internal structure of compact objects. Conversely, in the wave limit, where $GM\omega \ll 1$ the finite size effects of the compact object will be important. We will focus on the

latter regime. Moreover, the gravitational Raman amplitude can be separated into two parts. The first part captures the scattering against the background metric



(7.4.13)

which is made up of scattering with various mass monopole insertions on the worldline. The second part is the scattering against the star surface, which includes the tidal response



(7.4.14)

We parametrize our tidal response function by performing the low-frequency expansion

$$\langle Q_{L_1} Q_{L_2} \rangle_{\text{ret}}(\omega) = \delta_{\langle L_1 \rangle, \langle L_2 \rangle} \left(C_{\ell,0} + i\omega C_{\ell,1} + (i\omega)^2 C_{\ell,2} + \dots \right) . \tag{7.4.15}$$

Detailed calculations reveal that this amplitude exhibits both IR and UV divergences. IR divergences arise from the long-range nature of the Newtonian potential GM/r, analogous to Coulomb interactions discussed in Sect. 7.2. These diver-

gences will exponentiate when we resum all ladder-like diagrams as shown in Eq. (7.2.7). Regarding the UV divergences, they lead to renormalization group (RG) runnings in the tidal response function. Specifically focusing on scalar perturbations, the first type of RG running is induced by tidal effects. At the two-loop level, we derive the following RG equation:

$$\frac{\mathrm{d}\langle QQ\rangle_\ell}{\mathrm{d}\log\mu} = -(2GM\omega)^2\left[\frac{-11+15\ell(1+\ell)}{(-1+2\ell)(1+2\ell)(3+2\ell)}\langle QQ\rangle_\ell\right], \tag{7.4.16}$$

where μ is the renormalization scale. As we can see this RG equation depends on the specific form of the tidal response function. Remarkably, there is another type of running which is independent of the nature of the compact object. For the $\ell = 0$ sector, we find that

$$\frac{\mathrm{d}\langle QQ\rangle_{\ell=0}}{\mathrm{d}\log\mu} = (\text{self-induced}) - 4\pi(2GM)^3\,. \tag{7.4.17}$$

The second term is universal for all types of compact objects as it depends solely on their masses. For black holes, we can determine various coefficients in Eq. (7.4.15) by matching our EFT amplitude to black hole perturbation theory using the $\overline{\text{MS}}$ renormalization scheme.

$$C_{\ell=0,0} = 0, \qquad C_{\ell=1,0} = 0, \tag{7.4.18}$$

$$C_{\ell=0,2} = -4\pi r_s^3\left[\frac{1}{4\epsilon_{\text{UV}}} + \log(\mu r_s) + \frac{19}{12} + \gamma_E\right]. \tag{7.4.19}$$

We find that the static tidal Love numbers for $\ell = 0$ and $\ell = 1$ vanish, and the dynamical tidal Love numbers receive logarithmic corrections.

In summary, the worldline EFT offers a model-independent framework for studying the tidal effects of astrophysical compact objects. Given the diverse applications discussed in this section, the study of tidal effects is poised to enter a new era of precision science.

7.5 Radiation in Gravitational Observables

Holmfridur Sigridar Hannesdottir

QFT provides the framework to compute many different observables in particle physics and gravity. In previous sections, we have already discussed one such observable, the *scattering amplitude*, which encodes the probability amplitude for a set of "in" states to turn into a set of "out" states. In addition to the applications mentioned in previous sections, scattering amplitudes form the basis for computations in collider-physics experiments: we send in a pair of particles (such as protons at the Large Hadron Collider), and measure the outgoing particles – at least to the fullest extent possible with current technology.

However, many other observables are relevant in high-energy physics. In particular, recent years have seen the advent of measurements of *gravitational waves*, for example in the LIGO–Virgo–Kagra experiments. The physical setup in gravitational-wave measurements is different from the one in particle colliders. First, the measured waves originate in events occurring far away from Earth, such as black-hole merging, instead of a particle collision on Earth. Second, we only have access to the signal in a small corner of the phase space, and have to sum over unobserved configurations. This section is devoted to a discussion of the gravitational waveform as measured after a scattering of two black holes and is based on [31] [5]. For a more general discussion on different asymptotic observables and their analytic continuations, see the contribution [32].

Before we go on, let us emphasize that this waveform is different from the ones measured so far in the LIGO–Virgo–Kagra experiments. Current measurements correspond to radiation after *merging* of two heavy objects, while here we are computing the corresponding *scattering* events, that is, when the black holes do not form a bound state but rather scatter off each other on a hyperbolic orbit. When, and to which extent, future gravitational-wave experiments will be able to measure the scattering waveforms is an open question. One of the challenges in such measurements is that the scattering waveforms do not have the characteristic periodic behavior of the bound waveforms: as discussed in Sect. 7.2.3, the period of the merging waveform gets traded for a time delay of the scattering waveform. Nevertheless, it remains an open problem whether the bound-state waveforms can generally be obtained as analytic continuations of the scattering ones, see the discussion around (7.3.45) and Ref. [33].

7.5.1 The Gravitational Waveform

For a computation of the gravitational waveform, we model the black holes as heavy scalars with masses m_1 and m_2, and assume that they scatter off each other. Following the KMOC formalism [13], we write the incoming state as

$$|\psi\rangle_{\text{in}} = \prod_{i=1}^{2}\left[\int \mathrm{d}\Phi(p_i) f_i(p_i) e^{ib_i\cdot p_i/\hbar}\right]|p_1 p_2\rangle\,, \tag{7.5.1}$$

where $b \equiv b_1 - b_2$ is the impact parameter and the phase space is given by

$$\mathrm{d}\Phi(p_i) \equiv \frac{\mathrm{d}^{\mathrm{D}} p_i}{(2\pi)^{\mathrm{D}}} 2\pi\Theta(p^0)\delta(p_i^2 - m_i^2)\,, \tag{7.5.2}$$

for $i = 1, 2$.

[5] Compared to [31] and the contribution [32], we here use mostly minus signature as in the previous sections.

The functions $f(p_i)$ correspond to wavepackets, and we assume they are sharply peaked around the classical momenta of the black holes, as described in [13]. Following [34], we define the wavepackets as

$$\phi_b(p_1, p_2) \equiv e^{ip_1\cdot b_1/\hbar} e^{ip_2\cdot b_2/\hbar} f_1(p_1) f_2(p_2)\,. \tag{7.5.3}$$

We have labeled the state with the two black holes using their four-momenta p_1 and p_2, but we recall that they are on-shell external states, so $p_i^2 = m_i^2$. We define the state $|p_1 p_2\rangle$ as usual, with creation operators acting on the vacuum:

$$|p_1 p_2\rangle = a_1^\dagger a_2^\dagger |0\rangle\,. \tag{7.5.4}$$

The expectation value of the curvature in gravity is given by

$$R_{\mu\nu\rho\sigma}(x) = {}_{\text{in}}\langle\psi|S^\dagger \mathbb{R}_{\mu\nu\rho\sigma} S|\psi\rangle_{\text{in}}\,. \tag{7.5.5}$$

Expanding $\mathbb{R}_{\mu\nu\rho\sigma}$ out in its Fourier modes gives [34, 35]

$$R_{\mu\nu\rho\sigma}(x) = \kappa\,\text{Re}\sum_\eta \prod_{i\in\{1,2,1',2'\}} \left[\int \text{d}\Phi(p_i)\right] \phi_b(p_1, p_2)\phi_b^*(p_{1'}, p_{2'}) \tag{7.5.6}$$
$$\times \int \text{d}\Phi(k) k_{[\mu}\epsilon^\eta_{\nu]}(k) k_{[\rho}\epsilon^\eta_{\sigma]}(k) \langle p_1' p_2'|b_\eta(k)|p_1 p_2\rangle e^{-ik\cdot x}\,,$$

with $\kappa = \sqrt{32\pi G}$, and $b_\eta(k)$ is the future annihilation operator for a graviton with momentum k and helicity η.[6] We will henceforth suppress helicity labels for simplicity. The intuition behind this expression is that, since $\mathbb{R}_{\mu\nu\rho\sigma}$ depends linearly on the asymptotic-future creation and annihilation operators $b^\dagger(k)$ and $b(k)$, the expectation value of $b(k)$ emerges as the key dynamical quantity to compute. In this subsection, we will focus on computing the momentum-space expectation value

$$\text{Exp}_k \equiv \langle p_1' p_2'|b(k)|p_1 p_2\rangle\,. \tag{7.5.7}$$

The position-space waveform can then be obtained from (7.5.6), see [34, 36–39].

As presented in [32, Sec. 1], the future operators in the Heisenberg picture are related to the past operators via the S-matrix, which acts as an evolution operator:

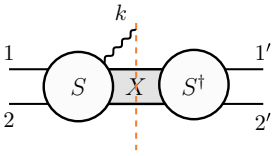
$$b(k) = S^\dagger a(k) S\,, \qquad b^\dagger(k) = S^\dagger a^\dagger(k) S\,, \tag{7.5.8}$$

[6] The creation and annihilation operators are often referred to as $a^\dagger_{\text{out}}$ and a_{out} in the literature, but we refer to them as b's instead to avoid clutter, as in [32, Sec. 1]. Similarly, we simply write a and $a^\dagger$ for the past operators, instead of a_{in} and $a^\dagger_{\text{in}}$.

where we assume that S is unitary, $SS^\dagger = \mathbb{1}$. Then, Exp_k becomes

$$\mathrm{Exp}_k = \langle p_1' p_2' | S^\dagger a(k) S | p_1 p_2 \rangle = \langle p_1' p_2' | S^\dagger \sum\!\!\!\!\!\!\int_X |X\rangle\langle Xk|S|p_1 p_2\rangle \,, \tag{7.5.9}$$

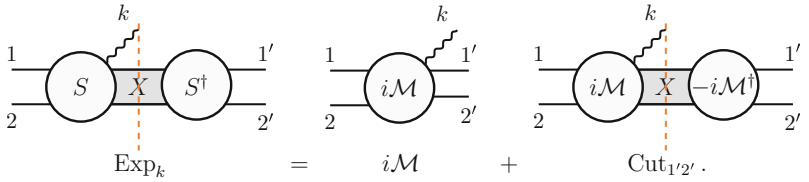
where we have inserted a complete set of states, $\mathbb{1} = \sum\!\!\!\!\!\!\int_X |X\rangle\langle X|$ and used that $\langle X|a(k) = \langle Xk|$. We call this expectation value the *waveshape* to avoid confusion with the position-space waveform. The waveshape is a product of an S-matrix for $p_1 p_2 \to kX$, and a conjugated S-matrix for $X \to p_{1'} p_{2'}$, summed and integrated over all states X, which we denote pictorially as

$\mathrm{Exp}_k =$ 

(7.5.10)

The shaded X in this picture corresponds to an insertion of a complete set of on-shell states with positive energy flowing across the dashed cut.

Using the definition of the amplitude $\mathcal{M}$ through $S = \mathbb{1} + i(2\pi)^{\mathrm{D}}\delta^{\mathrm{D}}(\sum p_i)\mathcal{M}$, we can rewrite the waveshape in terms of amplitudes and cut amplitudes,



(7.5.11)

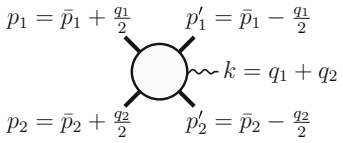
Importantly, the first term on the left-hand side is called *superclassical* (or *hyperclassical*), since it is proportional to a power of $\frac{1}{\hbar}$ in perturbation theory. As discussed in previous subsections, it thus does not make sense to take a classical limit $\hbar \to 0$ directly at the level of scattering amplitudes at a fixed order in perturbation theory.

Nevertheless, this power counting in $\hbar$ is entirely expected from classical physics: The exclusive amplitude in gravity is exponentially suppressed since the probability to create some fixed state is exponentially small. The amplitude therefore behaves as $\sim \mathrm{e}^{i\mathcal{S}/\hbar}$, where $\mathcal{S}$ is the action, and expanding out in Newton's constant G results in inverse powers of $\hbar$. The waveshape, Exp_k, is, on the other hand, a perfectly sensible classical observable since it sums over unobserved configurations. At a mathematical level, the cut term labeled $\mathrm{Cut}_{1'2'}$ in (7.5.11) precisely works to cancel off the $\frac{1}{\hbar}$ dependence of the scattering amplitude term, rendering Exp_k well defined in the $\hbar \to 0$ limit. In addition to canceling the superclassical contribution, the cut

term contributes to the infrared divergence and the finite part of Exp_k as we discuss below.

7.5.2 The Classical Limit

Our strategy will be to compute Exp_k in the classical limit of quantum field theory, using the KMOC formalism [13]. The waveshape Exp_k for measuring a graviton in the background of black-hole scattering is a five-point process, which we label with



(7.5.12)

We define the barred masses as $\bar{m}_i = \bar{p}_i^2$. The classical limit is obtained by taking the masses $\bar{m}_i$ of the black holes to be much larger than the Planck scale, and the impact parameter to be much larger than the wavelength of the graviton,

$$\bar{m}_1, \bar{m}_2 \gg M_{\text{pl}}\,, \qquad b \gg \lambda\,. \tag{7.5.13}$$

In the classification of (7.1.2), this condition corresponds to the Eikonal regime. This limit enforces the graviton to be soft, so

$$q_1 \sim \hbar\,, \qquad q_2 \sim \hbar\,, \qquad k \sim \hbar\,. \tag{7.5.14}$$

The coupling scales as $\kappa \sim \frac{1}{\sqrt{\hbar}}$, so Newton's constant scales as $G \sim \frac{1}{\hbar}$.

Note that this power counting is very different from the one used when doing computations purely within classical general relativity (GR), where $\hbar$ never makes an appearance. In GR, the perturbation series often involves a low-velocity expansion, in terms of powers of v^2, referred to as the *Post-Newtonian* (PN) *expansion*. When computing classical observables from QFT computations, the perturbation series is instead organized as a series expansion in G, but is exact in v, called the *Post-Minkowskian* (PM) *expansion*.

	0PN	1PN	2PN	3PN
1PM	G	v^2G	v^4G	v^6G
2PM	•	G^2	v^2G^2	v^4G^2
3PM	•	•	G^3	v^2G^3

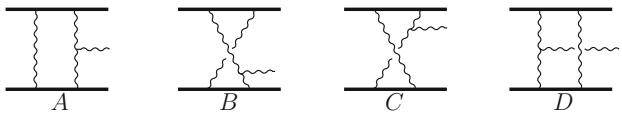
The 0PM expansion corresponds to flat space (no gravity), and the 0PN expansion corresponds to Newton's law of gravity.

7.5.3 One-Loop Contributions to the Waveshape

We now move onto analyzing the causal properties of the one-loop contributions to the waveshape. This subsection can be skipped if one wants to avoid the technical details.

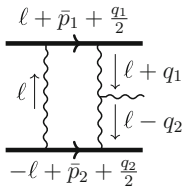
Using tensor reduction and integration by parts, we can reduce the set of integrals needed for the one-loop computation in the eikonal limit into a basis of 16 master integrals, which are subtopologies of a pentagon diagram along with the permutations of where the graviton legs attach. For illustration and simplicity, we focus here on the pentagon topology. We refer to Refs. [31, 34, 36–38] for the computation of all master integrals.

The four topologies we consider, which we label with A, B, C and D are



(7.5.15)

Using the following momentum labeling for the internal edges



(7.5.16)

the master integrals all belong to the following family of integrals,

$$G_{\mathcal{S}_i} = \mathrm{e}^{\epsilon\gamma_{\mathrm{E}}} \int \frac{\mathrm{d}^{\mathrm{D}}\ell}{i\pi^{\mathrm{D}/2}} \frac{1}{[\ell^2]^{a_1}[2\ell\cdot\bar{p}_1]^{a_2}[(\ell+q_1)^2]^{a_3}[(\ell-q_2)^2]^{a_4}[-2\ell\cdot\bar{p}_2]^{a_5}} \,. \tag{7.5.17}$$

where $\mathcal{S}_i = \{a_1, a_2, a_3, a_4, a_5\}$ labels a set of indices. For the pentagon diagram itself, we take $a_i = 1$ for all i. To write this integral family, we have used the expansion of the propagators appearing in the diagram from (7.5.16) in the eikonal limit, e.g.,

$$\frac{1}{[(-\ell+\bar{p}_2+\frac{1}{2}q_2)^2 - m_2^2]} = \frac{1}{\bar{m}_2}\frac{1}{[-2\ell\cdot\bar{v}_2]} - \frac{1}{\bar{m}_2^2}\frac{\ell\cdot(\ell-q_2)}{[-2\ell\cdot\bar{v}_2]^2} + \dots \,, \tag{7.5.18}$$

where we defined the velocities $\bar{v}_i = \bar{p}_i/\bar{m}_i$. We deliberately used square brackets to emphasize that the eikonal expansion holds for any $i\varepsilon$ prescription, which can be inserted into the brackets.

To compute the waveshape Exp_k, we have to sum over the contributions corresponding to diagrams A, B, C and D from (7.5.15) (including the relevant $+i\varepsilon$'s in all propagators), and, additionally, the cut of diagram D: A cut through the massive particles in the D topology allows for the graviton to be emitted before the cut, so this term must be included by (7.5.11). In the eikonal limit, the on-shell delta function of the cut propagator in topology D admits the following expansion,

$$\delta\Big[(\ell+\bar{p}_2-\tfrac{1}{2}q_2)^2-m_2^2\Big]=\frac{\delta\left(2\ell\cdot\bar{v}_2\right)}{\bar{m}_2}+\frac{\ell\cdot(\ell-q_2)}{\bar{m}_2^2}\delta'\left(2\ell\cdot\bar{v}_2\right)+\mathcal{O}\left(\bar{m}_2^{-3}\right), \tag{7.5.19}$$

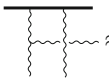
which is the analog of (7.5.18) for the propagators.

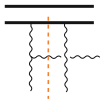
Looking at the contribution to the master integral (7.5.17) with $a_i = 1$ for all i, we get the following expansions corresponding to the pentagon-topology order in $\frac{1}{\bar{m}_2}$,

$I^A \equiv$  $\approx \int \mathrm{d}\hat{I}\frac{1}{(2\ell\cdot\bar{v}_1+i\varepsilon)(-2\ell\cdot\bar{v}_2+i\varepsilon)}\left[1-\frac{1}{\bar{m}_2}\frac{\ell\cdot(\ell-q_2)}{(-2\ell\cdot\bar{v}_2+i\varepsilon)}\right],$ (7.5.20a)

$I^B \equiv$  $\approx \int \mathrm{d}\hat{I}\frac{1}{(-2\ell\cdot\bar{v}_1+i\varepsilon)(-2\ell\cdot\bar{v}_2+i\varepsilon)}\left[1-\frac{1}{\bar{m}_2}\frac{\ell\cdot(\ell-q_2)}{(-2\ell\cdot\bar{v}_2+i\varepsilon)}\right],$ (7.5.20b)

$I^C \equiv$  $\approx \int \mathrm{d}\hat{I}\frac{1}{(2\ell\cdot\bar{v}_1+i\varepsilon)(2\ell\cdot\bar{v}_2+i\varepsilon)}\left[1-\frac{1}{\bar{m}_2}\frac{\ell\cdot(\ell-q_2)}{(2\ell\cdot\bar{v}_2+i\varepsilon)}\right],$ (7.5.20c)

$I^D \equiv$  $\approx \int \mathrm{d}\hat{I}\frac{1}{(-2\ell\cdot\bar{v}_1+i\varepsilon)(2\ell\cdot\bar{v}_2+i\varepsilon)}\left[1-\frac{1}{\bar{m}_2}\frac{\ell\cdot(\ell-q_2)}{(2\ell\cdot\bar{v}_2+i\varepsilon)}\right],$ (7.5.20d)

$\mathrm{Cut}I^D \equiv$  $\approx 4\pi^2\int \mathrm{d}\hat{I}\,\delta(2\ell\cdot\bar{v}_1)\left[\delta(2\ell\cdot\bar{v}_2)+\frac{\ell\cdot(\ell-q_2)}{\bar{m}_2}\delta'(2\ell\cdot\bar{v}_2)\right],$ (7.5.20e)

where we have labelled $\mathcal{S}_1 \equiv \{1, 1, 1, 1, 1\}$, and used a short-hand notation for the phase space and the graviton propagators,

$$\int \mathrm{d}\hat{I} \equiv \frac{\mathrm{e}^{\epsilon\gamma_{\mathrm{E}}}}{\bar{m}_1\bar{m}_2} \int \frac{\mathrm{d}^{\mathrm{D}}\ell}{i\pi^{\mathrm{D}/2}[\ell^2]^{a_1}[(\ell+q_1)^2]^{a_3}[(\ell-q_2)^2]^{a_4}} \,. \tag{7.5.21}$$

Using the distributional identity $\frac{1}{x+i\varepsilon} - \frac{1}{x-i\varepsilon} = -2\pi i\delta(x)$, we see that when adding these contributions (i.e. computing $I^A + I^B + I^C + I^D + \mathrm{Cut}I^D$), the leading term in $\frac{1}{\bar{m}_2}$ cancels.

To analyze the subleading terms, let us take a step back and analyze more generally different ways in which these diagrams could be added or subtracted to contribute to the waveform. First, note that the topologies B, C and D are permutations of A, obtained by taking $\bar{p}_i \to -\bar{p}_i$ in different ways. Generally, we can write,

$$\begin{aligned}\mathrm{Exp}_k = \sum_i \Big[& c_i(\bar{p}_1, \bar{p}_2) I^A_{\mathcal{S}_i} + c_i(-\bar{p}_1, \bar{p}_2) I^B_{\mathcal{S}_i} \\ & + c_i(\bar{p}_1, -\bar{p}_2) I^C_{\mathcal{S}_i} + c_i(-\bar{p}_1, -\bar{p}_2) \left(I^D_{\mathcal{S}_i} + \mathrm{Cut}\, I^D_{\mathcal{S}_i} \right) \Big] ,\end{aligned} \tag{7.5.22}$$

where each $\mathcal{S}_i$ is a set of indices, and the sum is over all master integrals needed to compute Exp_k. The numerator factors c_i are those obtained by tensor reductions and integration by parts. Recall that the numerators come from vertices between the heavy black holes and the gravitons,

$$\underbrace{\qquad\qquad}_{\sim\sim} \propto \sqrt{G}\bar{m}_i^2 \,, \tag{7.5.23}$$

so the leading term in the classical expansion of the numerator of the pentagon diagrams will be of order $\bar{m}_1^4\bar{m}_2^4$. The first subleading order will come from numerators that are proportional to either $\bar{m}_1^3\bar{m}_2^4$ or $\bar{m}_1^4\bar{m}_2^3$.

The numerator terms that are proportional to $\bar{m}_1^4\bar{m}_2^4$ are symmetric in $\bar{p}_i \to -\bar{p}_i$ for $i = 1, 2$, meaning that, according to (7.5.22), the five terms from (7.5.20) must be added up to form the waveshape. As noted before, we see that the first terms in the square brackets all cancel each other, i.e., the term proportional to $\bar{m}_1^4\bar{m}_2^4$ vanishes. According to the power counting, these leading terms that cancel are precisely the superclassical contributions that scale as $\frac{1}{\hbar}$ around $\mathrm{D} = 4$ spacetime dimensions. However, this symmetric contribution also contributes to the next order,

$$\begin{aligned}&\bar{m}_1^4\bar{m}_2^4(I^A + I^B + I^C + I^D + \mathrm{Cut}\, I^D) \\ &\quad = \bar{m}_1^4\bar{m}_2^3 \int \mathrm{d}\hat{I}\, \delta(2\ell\cdot\bar{v}_1) \frac{4\pi i\, \ell\cdot(\ell - q_2)}{(-2\ell\cdot\bar{v}_2 + i\varepsilon)^2} + \ldots ,\end{aligned} \tag{7.5.24}$$

where the dots indicate the terms that are of $\mathcal{O}(\bar{m}_1^2\bar{m}_2^3)$ (recall the factors of $\bar{m}_i$ in $\mathrm{d}\hat{I}$), as well as the terms of higher orders in $1/\bar{m}_i$, which are suppressed in the classical expansion. Similarly, the subleading numerator factors of order $\bar{m}_1^3\bar{m}_2^2$ are antisymmetric under $\bar{p}_2 \to -\bar{p}_2$, so topologies C and D get subtracted in their contribution to Exp_k. These terms, that have antisymmetric numerators, contribute as

$$\bar{m}_1^4\bar{m}_2^3(I^A+I^B-I^C-I^D+\mathrm{Cut}\, I^D) = \bar{m}_1^4\bar{m}_2^3 \int \mathrm{d}\hat{I}\, \delta(2\ell\cdot\bar{v}_1)\frac{-4\pi i}{-2\ell\cdot\bar{v}_2+i\varepsilon} + \dots\,, \tag{7.5.25}$$

where, again, the dots indicate terms of $\mathcal{O}(\bar{m}_1^2\bar{m}_2^3)$ or higher orders in $1/\bar{m}_i$.

As a last remark about this computation, note that the role of the cut contribution $\mathrm{Cut}I^D$ is not solely to cancel the classical term of Exp_k, but also to change the causal properties of the propagators in the diagrams. In particular, if we had computed the corresponding amplitude contribution $\mathcal{M}$, the propagators would have been symmetric, principal-valued propagators, see [31,34,36–38]. However, after adding $\mathrm{Cut}I^D$, we get retarded propagators instead, as indicated with the $i\varepsilon$ prescription in (7.5.25) and (7.5.24). The physical interpretation is that these propagators enforce the causal condition of the waveshape Exp_k: the emitted graviton is in the future of all other particles.

7.5.4 Infrared Divergence and Physical Interpretation

We have already seen two ways in which the cut term on the right-hand side of (7.5.11) contributes to the waveshape. First, it eliminates the superclassical $\frac{1}{\hbar}$ contribution, and second, it modifies the Feynman propagators to become retarded propagators, thus enforcing that the graviton is in the future of the incoming black holes. The cut term has yet another role, however; it enforces the infrared divergence to match with the classical prediction for the Shapiro time delay of this process.

Let us first discuss what result we expect for the infrared divergence for the scattering amplitude $\mathcal{M}$ and for the waveshape Exp_k. It is well-known that the infrared divergence of the scattering amplitude $\mathcal{M}$ is caused by the Shapiro time delay for the graviton escaping the black hole potential, causing its wavelength to getting redshifted. We denote this Shapiro time delay as Δt_{obs}. Its value in classical GR is

$$\Delta t_{\mathrm{obs}} = 2G\,\hat{k}\cdot(\bar{p}_1+\bar{p}_2)\log\frac{r_{\mathrm{obs}}}{b}\,, \tag{7.5.26}$$

where b is the impact parameter, r_{obs} is the distance between the observation point and the black holes, and $\hat{k}$ is the unit vector in the direction of the graviton. Note that this time delay is not present in electromagnetism, since photons do not get delayed or redshifted in the presence of an electromagnetic potential. Since the time delay is

logarithmically divergent as $r_{\rm obs} \to \infty$, we expect the avatar of this time delay to be an infrared divergence in the perturbation theory computation, where the framework enforces $r_{\rm obs} \to \infty$ from the outset. The infrared divergences can be computed using Weinberg soft factors, resulting in the following one-loop divergence,

$$\mathcal{M}^{(1\text{-loop})}\Big|_{\text{IR div}} = \mathcal{M}^{(\text{tree})} \times 2iG\,k\cdot(\bar{p}_1+\bar{p}_2)\log\frac{\Lambda}{\mu_{\rm IR}}\,, \tag{7.5.27}$$

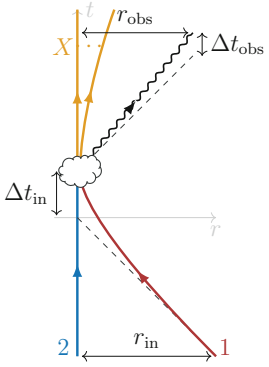
where $\Lambda \sim b^{-1}$ is a physical momentum scale, and $\mu_{\rm IR}$ is an infrared renormalization scale. When including all-orders contributions, these infrared divergences exponentiate to recover the classical time delay.

The waveshape Exp_k contains an additional source of Shapiro time delay which is not present in the exclusive amplitude $\mathcal{M}$. Since the waveshape is an "in-in" observable, see [40], there is an additional time delay or advance coming from fixing the initial distance $r_{\rm in}$ between the two black holes. This contribution also has a classical explanation: Imagine two black holes venturing into a scattering process with momenta p_1 and p_2. If they are moving at non-relativistic speeds, they will start attracting each other and scatter off each other at an earlier time than expected if they had been moving on free trajectories. In fact, the time advance is logarithmic in the initial separation between them, $r_{\rm in}$. This attraction results in a time advance of the scattering point compared with free motion. It turns out that relativistically, there is a value of $\bar{p}_1\cdot\bar{p}_2$ at which the time advance turns into a time delay instead. In any case, the effect can easily be computed in classical GR assuming black holes moving in a Schwarzschild metric [31]: The time delay due to the fixed initial distance between the two black holes is

$$\Delta t_{\rm in} = G\,\hat{k}\cdot(\bar{p}_1+\bar{p}_2)\times\frac{2-3\bar{y}^{-2}}{(1-\bar{y}^{-2})^{\frac{3}{2}}}\log\frac{r_{\rm in}}{b}\,, \tag{7.5.28}$$

where $\bar{y} = \bar{v}_1\cdot\bar{v}_2$. Note that $\Delta t_{\rm in} < 0$ corresponds to a time advance.

Pictorially, we can represent the contributions from the two sources of time delay for Exp_k in gravity as



(7.5.29)

Adding the two time delays, from (7.5.26) and (7.5.28), we find that the corresponding IR divergence of Exp_k when computed at one loop in the QFT is

$$\underbrace{\text{Exp}_k^{(1\text{-loop})}\Big|_{\text{div}} = \text{Exp}_k^{(\text{tree})} \times iGk\cdot(\bar{p}_1+\bar{p}_2)\left(2+\frac{2-3\bar{y}^{-2}}{(1-\bar{y}^{-2})^{\frac{3}{2}}}\right)\log\frac{\Lambda}{\mu_{\text{IR}}}}_{\text{(classical prediction)}}, \tag{7.5.30}$$

where, as before $\Lambda \sim b^{-1}$ is a physical scale and μ_{IR} is an infrared scale. This IR divergence was computed in [31], and was later confirmed in explicit computations of Exp_k at one loop in [34, 36, 37, 39, 41–43].

Note that alternatively, we can compute the leading IR divergence using Wilson lines, analogous to the QED computation in [32, Sec. 2.3]. The difference between QED and gravity in this case is that here we will get an extra contribution from (7.5.26), which corresponds to the time delay that the graviton experiences when moving in the field of the two massive particles. The computation in gravity was carried out in [31].

7.6 Hierarchical 3-Body Problem

Anna M. Wolz

In this section, we will complicate the already complicated story of 2-body scattering by adding an extra body. So, we will have three compact massive objects in total, interacting via gravity.

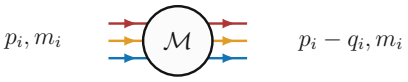
The 3-body problem is a notoriously difficult problem in astrophysics. It was approached from a scattering amplitudes and EFT perspective in Ref. [44], whose construction we will review in this section. Then, following [45], we will motivate a version of the 3-body problem in which one of the massive objects is very far away from the other two, called the *hierarchical limit*.

In this formalism, one can treat both bound and unbound states. For example, we can consider the case in which a supermassive black hole interacts with two other heavy bodies, out of which one falls in and the other flies out. This setup is relevant to many astrophysical systems. If we assume that velocities of the black holes are small compared with the speed of light, we can use the PN expansion in the bound case. Otherwise, we need the PM expansion.

Of course, 3-body systems are extremely complicated even in Newtonian gravity. Newtonian solutions to the equations of motion are unknown in most situations and otherwise difficult to obtain. Within the framework of GR, it is difficult to even explicitly write down all the interactions, since we have intrinsic N-body effects. Due to this complexity in GR, we will instead approach the 3-body problem using the tools from particle physics, similar to the ones used in the previous sections. We will be able to make headway by using approximations and treating computations perturbatively.

As we mentioned above, the main tools at our disposal are the PN expansion, which works well for bound systems that move slowly and where the bodies are far apart (equivalently, we deal with weak fields so Newton's constant G is small). In this approximation, the expansion parameter is Gv^2, where v denotes the velocities of the bodies. On the other hand, PM is relevant to unbound encounters, where the bodies move fast, though they are still far apart. Then, we expand in G alone. Amplitudes are most naturally suited for the PM case where we have relativistic high-energy scattering, and perturbation theory is already linked to the G-expansion.

The setup will be as follows. We consider $3 \to 3$ scattering of masses m_i, whose initial momenta are p_i and the final momenta are $p_i - q_i$ for $i = 1, 2, 3$:



(7.6.1)

The initial momentum of each body can be written as $p_i^\mu = (E_i(\mathbf{p}_i), \mathbf{p}_i)$, and the momentum transfer of each body is $q_i^\mu = (q_i^0, \mathbf{q}_i)$. We can go freely between the positions of the objects $\mathbf{r}_i$ and the spatial momentum transfers $\mathbf{q}_i$ by performing a Fourier transform

$$\mathbf{r}_i \overset{\text{FT}}{\longleftrightarrow} \mathbf{q}_i \,. \tag{7.6.2}$$

In this section, we will describe how to calculate the classical, conservative 3-body potential $V^{(3)}(\{\mathbf{p}, \mathbf{q}\})$ (from now on, curly brackets denote the dependence on the momenta $\mathbf{p}_i$ and momentum transfers $\mathbf{q}_i$ for all $i = 1, 2, 3$) in momentum space, as done in [44]. The Fourier transform of this function into position space $V^{(3)}(\{\mathbf{p}, \mathbf{r}\})$ in the hierarchical limit is described in [45].

The way we will calculate $V^{(3)}(\{\mathbf{p}, \mathbf{q}\})$ below is by taking the classical limit of the amplitude in (7.6.1) with scalars minimally coupled to gravity. More precisely, we will compute the amplitude from a low-energy EFT of particles interacting via the potential $V(\{\mathbf{p}, \mathbf{q}\})$ and then formally match it onto the amplitude $\mathcal{M}$ using the Lippmann–Schwinger equation.

Note that the 3-body problem necessarily involves potentials with two types of terms:

$$V = \sum_{\substack{\text{pairs} \\ i,j}} V_{ij}^{(2)} + V^{(3)} \,, \tag{7.6.3}$$

where the first piece encodes the 2-body interactions and the second one is intrinsically 3-body and it depends on all three momenta $\mathbf{p}_i$ and positions $\mathbf{r}_i$. Note that we do not have the 3-body potential $V^{(3)}$ in Newtonian dynamics.

Here we are interested in the leading $V^{(3)}$ interaction. In this case, leading order translates to $\mathcal{O}(G^2)$, i.e., the 2PM order.

7.6.1 Lippmann–Schwinger Equation

The setup looks simple so let us get to it. We start with the Einstein–Hilbert action minimally coupled to scalars:

$$S = \int \mathrm{d}^4x\sqrt{-g}\left(\frac{-R}{16\pi G} + \sum_{i=1}^{3}\left((\nabla_\mu\phi_i)^2 - m_i^2\phi_i^2\right)\right), \tag{7.6.4}$$

where ϕ is a real field. One could of course add higher-derivative corrections, but here we are working in the simplest setup so we approximate black holes as point particles. This approximation is valid when their Schwarzschild radii R_S are much smaller than the distances $|\mathbf{r}_i - \mathbf{r}_j|$ between the black holes. If we wanted to include finite-size effects, charge, etc., we could use exactly the same terms as in the 2-body case.

In order to obtain the 3-body potential at 2PM, the necessary ingredients will be the tree-level matrix elements

$$\mathcal{M}_G^{(2)} \qquad \text{and} \qquad \mathcal{M}_{G^2}^{(3)} \tag{7.6.5}$$

for the $2 \to 2$ and $3 \to 3$ cases respectively. The subscript denotes the order in G at which are working in. We need them computed in a general frame, which is a slight complication compared to the usual 2-body case which are usually calculated in the center-of-mass frame.

The classical limit for us means large impact parameter, or equivalently, small momentum transfer q. For the $2 \to 2$ case, the leading scaling in q will be $\mathcal{O}(q^{-2})$ and for $3 \to 3$, it will be $\mathcal{O}(q^{-4})$.

Using Feynman rules, the two amplitudes at tree level are simply given by

$\mathcal{M}_G^{(2)} =$  (7.6.6)

and

$\mathcal{M}_{G^2}^{(3)} =$  (7.6.7)

We are keeping only the diagrams that contribute non-trivially at the orders we require. We use the notation $M^{(N)}$ to denote the $N \to N$ amplitude multiplied by a normalization factor:

$$M^{(N)} = \prod_{i=1}^{N} \frac{1}{2\sqrt{E_i(\mathbf{p}_i)E_i(\mathbf{p}_i - \mathbf{q}_i)}}\mathcal{M}^{(N)}. \tag{7.6.8}$$

Next, we will write down a low-energy EFT of relativistic particles. It is governed by the Hamiltonian

$$H = \sum_{i=1}^{3} \sqrt{\mathbf{p}_i^2 + m_i^2} + V(\{\mathbf{p}, \mathbf{r}\}) \,. \tag{7.6.9}$$

As mentioned above, the potential is related via a Fourier transform to $V(\{\mathbf{p}, \mathbf{q}\})$. The Lippmann–Schwinger equation can formally be expressed as:

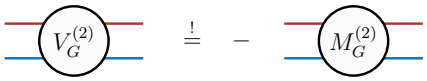
$$T(\{\mathbf{p}, \mathbf{q}\}) \stackrel{!}{=} -V(\{\mathbf{p}, \mathbf{q}\}) - \int_{\mathbf{k}} \frac{V(\{\mathbf{p}, \mathbf{p}-\mathbf{k}\})\, V(\{\mathbf{k}, \mathbf{k}-\mathbf{p}+\mathbf{q}\})}{\sum_{i=1}^{3}[E_i(\mathbf{p}) - E_i(\mathbf{k})] + i\varepsilon} + \dots \tag{7.6.10}$$

In this equation, T denotes the scattering amplitudes, i.e. the matrix elements of the interacting part of the S-matrix (as in $S = \mathbb{1} + i\hat{T}$). We have written the iteration terms and stopped at the quadratic order in V since this will be sufficient for our purposes. Note that $\stackrel{!}{=}$ denotes that the potential is defined only up to terms that vanish on-shell, which are the gauge/coordinate artifacts.

We are going to think of this equation in terms of scattering amplitudes. It is actually valid for any N-body scattering, but here we will apply it only to $N = 2, 3$ specifically. The potentials $V^{(2)}$ and $V^{(3)}$ are like the Wilson coefficients in the EFT. We will use Feynman diagrams to compute T and then match on to the right hand-side of the above equation. Note that T contains the fully connected term, but also partially disconnected pieces when $N \geqslant 3$.

7.6.2 Solving the 2-Body Potential

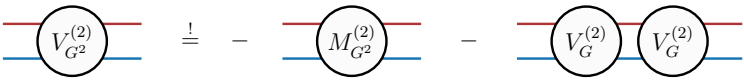
Let us first deal with the 2-body case, which we have already encountered in Sect. 7.3. At $\mathcal{O}(G)$, we have the tree-level contribution



(7.6.11)

The notation $V_G^{(2)}$ denotes the 2-body potential at $\mathcal{O}(G)$. This is simply the Newtonian potential, computed with the t-channel exchange of a graviton.

At the next order, $\mathcal{O}(G^2)$, we encounter the contribution from Feynman diagrams of $\mathcal{M}$ at this order, but also the iteration of the $V^{(2)}$ terms we computed above:



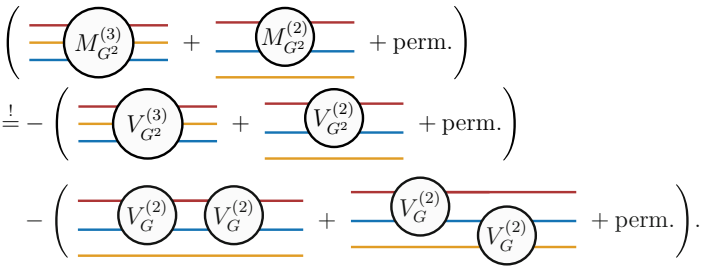
(7.6.12)

The notation of two lines between $V_G^{(2)}$ in the last term represents the integration over the appropriate phase space of the intermediate particles (as in the last term of

(7.6.10)). This procedure allows us to determine $V^{(2)}_{G^2}$ diagrammatically, which is all we will need.

7.6.3 Solving the 3-Body Potential

Let us move on to the 3-body case. Now, the T-matrix contains contributions from both $3 \to 3$ and $2 \to 2$ amplitudes. Diagrammatically, the expansion of (7.6.10) looks as follows:



(7.6.13)

The second line corresponds to the terms linear in V in the Lippmann–Schwinger equation and the third line involves the quadratic terms. Our goal is to isolate the piece $V^{(3)}_{G^2}$. We plug in (7.6.12) for $V^{(2)}_{G^2}$ with the spectator particle (in yellow). This cancels with the $M^{(2)}_{G^2}$ piece on the LHS and the "loop" iteration piece on the LHS of the third line. All of the disconnected pieces therefore cancel. Solving for $V^{(3)}_{G^2}$, the final expression reads diagrammatically

(7.6.14)

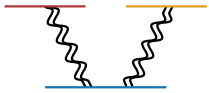
We now have all the ingredients to compute the effective 3-body potential. After the dust settles, the final answer is:

$$V^{(3)}_{G^2} \stackrel{!}{=} \sum_{(i,j,k)\in S_3} \left(-\frac{\mathcal{M}^{(3)}_{G^2,ijk}}{8E_iE_jE_k} - \frac{V^{(2)}_{G,ij}(\mathbf{p}_i,\mathbf{p}_j;\mathbf{q}_i)\, V^{(2)}_{G,kj}(\mathbf{p}_k,\mathbf{p}_j+\mathbf{q}_i;\mathbf{q}_k)}{E_j + q^0_i - \sqrt{(\mathbf{p}_j+\mathbf{q}_i)^2+m_j^2}} \right), \tag{7.6.15}$$

where S_3 is the set of distinct permutations of the particle labels $(1, 2, 3)$ and $\mathcal{M}^{(3)}$ is the fully connected matrix element for the $3 \to 3$ scattering process.

Let us do some cross checks on this result. There are a couple of criteria that a classical potential should satisfy. First, it should behave as $\mathcal{O}(q^{-4})$. Naively, it might appear from (7.6.15) that the three-body potential behaves as $\mathcal{O}(q^{-5})$, which

would lead to a superclassical $\sim 1/\hbar$ scaling. Second, the potential should have no matter singularities, i.e., only the graviton poles should survive. Let us illustrate these problems at the level of the individual diagram, say:



(7.6.16)

One can indeed show that it scales as $\mathcal{O}(q^{-5})$. Likewise, it contains an explicit matter pole (associated to the blue propagator) of the form

$$\frac{1}{(p_i + q_j)^2 - m_i^2} \,. \tag{7.6.17}$$

After some work, one can show that these two problems go away once we sum over all the diagrams and iteration terms. In other words, (7.6.15) goes as $\mathcal{O}(q^{-4})$ and does not have any matter poles, although it is not manifest at the level of the formula (7.6.15).

These cross-checks give us confidence that (7.6.15) is the correct 3-body potential at the leading order. Note that it is supposed to be valid for both unbound and bound orbits. The authors of [44] checked the result for $V^{(3)}_{G^2}(\{\mathbf{p}, \mathbf{q}\})$ with [46], who computed the 3-body effective potential using other techniques.

References

1. N. Arkani-Hamed, P. Benincasa, S. Caron-Huot, M. Correia, S. Curry, M. Giroux, F.M. Haehl, H.S. Hannesdottir, M.T. Hansen, G. Isabella, J. Lebl, M.H.G. Lee, S. Mizera, E. Pajer, C. Pasiecznik, M. Rangamani, F. Vazão, A.M. Wolz, Z. Zhou, Records from the S-Matrix Marathon: Selected Topics on Scattering Amplitudes. Springer Lecture Notes in Physics (2025)
2. LIGO Scientific, Virgo Collaboration, B.P. Abbott et al., Observation of gravitational waves from a binary black hole merger. Phys. Rev. Lett. **116**, 061102 (2016) [1602.03837]
3. I.T. Todorov, Quasipotential equation corresponding to the relativistic Eikonal approximation. Phys. Rev. D **3**, 2351 (1971)
4. V. Singh, Analyticity in the complex angular momentum plane of the Coulomb scattering amplitude. Phys. Rev. **127**, 632 (1962)
5. C. Cheung, I.Z. Rothstein, M.P. Solon, From scattering amplitudes to classical potentials in the post-Minkowskian expansion. Phys. Rev. Lett. **121**, 251101 (2018) [1808.02489]
6. G. Kälin, R.A. Porto, Post-Minkowskian effective field theory for conservative binary dynamics. J. High Energy Phys. **11**, 106 (2020) [2006.01184]
7. Z. Bern, J.P. Gatica, E. Herrmann, A. Luna, M. Zeng, Scalar QED as a toy model for higher-order effects in classical gravitational scattering. J. High Energy Phys. **08**, 131 (2022) [2112.12243]
8. B. Bellazzini, G. Isabella, M.M. Riva, Classical vs quantum eikonal scattering and its causal structure. J. High Energy Phys. **04**, 023 (2023) [2211.00085]
9. M. Correia, G. Isabella, The Born regime of gravitational amplitudes. 2406.13737

10. J.J. Sakurai, J. Napolitano, *Modern Quantum Mechanics*. Quantum Physics, Quantum Information and Quantum Computation (Cambridge University Press, Cambridge, 2020). https://doi.org/10.1017/9781108587280
11. A. Buonanno, T. Damour, Effective one-body approach to general relativistic two-body dynamics. Phys. Rev. D **59**, 084006 (1999) [gr-qc/9811091]
12. P.P. Fiziev, I.T. Todorov, Effective one-body approach to the relativistic two-body problem. Phys. Rev. D **63**, 104007 (2001) [gr-qc/0010104]
13. D.A. Kosower, B. Maybee, D. O'Connell, Amplitudes, observables, and classical scattering. J. High Energy Phys. **02**, 137 (2019) [1811.10950]
14. J.D. Jackson, *Classical Electrodynamics* (Wiley, New York, 1998)
15. W.D. Goldberger, I.Z. Rothstein, An effective field theory of gravity for extended objects. Phys. Rev. D **73**, 104029 (2006) [hep-th/0409156]
16. I.Z. Rothstein, Progress in effective field theory approach to the binary inspiral problem. Gen. Rel. Grav. **46**, 1726 (2014)
17. R.A. Porto, The effective field theorist's approach to gravitational dynamics. Phys. Rep. **633**, 1 (2016) [1601.04914]
18. W.D. Goldberger, Effective Field Theory for Compact Binary Dynamics. 2212.06677
19. E. Poisson, C.M. Will, *Gravity* (2014). https://doi.org/10.1017/CBO9781139507486
20. T. Pitre, E. Poisson, General relativistic dynamical tides in binary inspirals without modes. Phys. Rev. D **109**, 064004 (2024) [2311.04075]
21. S. Van Velzen, G.R. Farrar, S. Gezari, N. Morrell, D. Zaritsky, L. Östman, M. Smith, J. Gelfand, A.J. Drake, Optical discovery of probable stellar tidal disruption flares. Astrophys. J. **741**, 73 (2011)
22. B. Mockler, J. Guillochon, E. Ramirez-Ruiz, Weighing Black Holes using tidal disruption events. Astrophys. J. **872**, 151 (2019) [1801.08221]
23. E. Hammerstein et al., The final season reimagined: 30 tidal disruption events from the ZTF-I survey. Astrophys. J. **942**, 9 (2023) [2203.01461]
24. H.S. Chia, Z. Zhou, M.M. Ivanov, Bring the Heat: Tidal Heating Constraints for Black Holes and Exotic Compact Objects from the LIGO-Virgo-KAGRA Data. 2404.14641
25. M.M. Ivanov, Y.-Z. Li, J. Parra-Martinez, Z. Zhou, Gravitational Raman scattering in effective field theory: a scalar tidal matching at O(G3). Phys. Rev. Lett. **132**, 131401 (2024) [2401.08752]
26. S.A. Teukolsky, Perturbations of a rotating black hole. 1. Fundamental equations for gravitational electromagnetic and neutrino field perturbations. Astrophys. J. **185**, 635 (1973)
27. G. Aminov, A. Grassi, Y. Hatsuda, Black Hole quasinormal modes and Seiberg–Witten theory. Ann. Henri Poincare **23**, 1951 (2022) [2006.06111]
28. G. Bonelli, C. Iossa, D.P. Lichtig, A. Tanzini, Exact solution of Kerr black hole perturbations via CFT2 and instanton counting: Greybody factor, quasinormal modes, and Love numbers. Phys. Rev. D **105**, 044047 (2022) [2105.04483]
29. G. Bonelli, C. Iossa, D. Panea Lichtig, A. Tanzini, Irregular Liouville correlators and connection formulae for Heun functions. Commun. Math. Phys. **397**, 635 (2023) [2201.04491]
30. Y.F. Bautista, G. Bonelli, C. Iossa, A. Tanzini, Z. Zhou, Black hole perturbation theory meets CFT2: Kerr-Compton amplitudes from Nekrasov-Shatashvili functions. Phys. Rev. D **109**, 084071 (2024) [2312.05965]
31. S. Caron-Huot, M. Giroux, H.S. Hannesdottir, S. Mizera, What can be measured asymptotically? J. High Energy Phys. **01**, 139 (2024) [2308.02125]
32. S. Caron-Huot, M. Giroux, H.S. Hannesdottir, S. Mizera, C. Pasiecznik, Records from the S-matrix marathon: asymptotic observables. [2502.13021]
33. T. Adamo, R. Gonzo, A. Ilderton, Gravitational Bound Waveforms from Amplitudes. 2402.00124
34. A. Elkhidir, D. O'Connell, M. Sergola, I.A. Vazquez-Holm, Radiation and Reaction at One Loop. 2303.06211v1
35. A. Cristofoli, R. Gonzo, D.A. Kosower, D. O'Connell, Waveforms from amplitudes. Phys. Rev. D **106**, 056007 (2022) [2107.10193]

36. A. Brandhuber, G.R. Brown, G. Chen, S. De Angelis, J. Gowdy, G. Travaglini, One-loop Gravitational Bremsstrahlung and Waveforms from a Heavy-Mass Effective Field Theory. 2303.06111v3
37. A. Herderschee, R. Roiban, F. Teng, The Sub-Leading Scattering Waveform from Amplitudes. 2303.06112v2
38. A. Georgoudis, C. Heissenberg, I. Vazquez-Holm, Inelastic Exponentiation and Classical Gravitational Scattering at One Loop. 2303.07006v1
39. A. Georgoudis, C. Heissenberg, I. Vazquez-Holm, Addendum to: Inelastic exponentiation and classical gravitational scattering at one loop. J. High Energy Phys. **2024**, 161 (2024) [2312.14710]
40. F.M. Haehl, M. Rangamani, Records from the S-Matrix Marathon: Schwinger-Keldysh Formalism (2024) 2410.10602
41. L. Bohnenblust, H. Ita, M. Kraus, J. Schlenk, Gravitational Bremsstrahlung in black-hole scattering at $\mathcal{O}\left(G^3\right)$: linear-in-spin effects. J. High Energy Phys. **11**, 109 (2024) [2312.14859]
42. D. Bini, T. Damour, A. Geralico, Comparing one-loop gravitational bremsstrahlung amplitudes to the multipolar-post-Minkowskian waveform. Phys. Rev. D **108**, 124052 (2023) [2309.14925]
43. D. Bini, T. Damour, S. De Angelis, A. Geralico, A. Herderschee, R. Roiban, F. Teng, Gravitational waveforms: a tale of two formalisms. Phys. Rev. D **109**, 125008 (2024) [2402.06604]
44. C.R.T. Jones, M. Solon, Scattering amplitudes and N-body post-Minkowskian Hamiltonians in general relativity and beyond. J. High Energy Phys. **02**, 105 (2023) [2208.02281]
45. M.P. Solon, A.M. Wolz, The Hierarchical Three-Body Problem at $\mathcal{O}(G^2)$. 2408.08850
46. F. Loebbert, J. Plefka, C. Shi, T. Wang, Three-body effective potential in general relativity at second post-Minkowskian order and resulting post-Newtonian contributions. Phys. Rev. D **103**, 064010 (2021) [2012.14224]