UC Santa Barbara
UC Santa Barbara Electronic Theses and Dissertations

Title

Optimal Transport for High Energy Physics

Permalink
https://escholarship.org/uc/item/93n2s2ff

Author
Cai, Tianji

Publication Date
2023

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

University of California
Santa Barbara

Optimal Transport for High Energy Physics

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Physics

by

Tianji Cai

Committee in charge:

Professor Nathaniel Craig, Chair
Professor Mark Srednicki
Professor Jeffrey Richman

June 2023

The Dissertation of Tianji Cai is approved.

Professor Mark Srednicki

Professor Jeffrey Richman

Professor Nathaniel Craig, Committee Chair

June 2023

Optimal Transport for High Energy Physics

Copyright © 2023

by

Tianji Cai

Acknowledgements

First and foremost, I would like to say a big thank you to my Ph.D. advisor, Nathaniel Craig, for always believing in me more than I believe in myself and for having provided me with all the great opportunities to prove and to surprise myself. I would not have the courage and determination to further pursue the academic path were it not for Nathaniel's caring support and the time and freedom he has given me to explore my research interests. Thank you, Nathaniel, for helping me grow into the researcher and the individual that I aspire to be!

My gratitude also goes without saying to my mentor on the math side, Katy Craig, for leading me into the beautiful world of optimal transport and for guiding my exploration of this enchanting land of interdisciplinary research. I enjoy every moment of our project meetings, our casual chats, and our brainstorming sessions. Such collaborative experience is invaluable to me and has shaped and fueled my commitment for more cross-disciplinary work. Research aside, I am especially grateful to Katy for all the helps and advices she offers me in life, and to both Nathaniel and Katy for fostering such a joyful and home-like work environment. This has really boosted my overall happiness while I am studying abroad, which in turn unlocks my research potential.

The work presented here would not be possible without the joint efforts and immense help from all my collaborators. From Bernhard Schmitzer and Matthew Thorpe, I have learnt a great deal of optimal transport from a mathematical point of view. As the project was conducted during the peak of the pandemic, I would especially like to thank them for their patience and kindness when I was in quarantine, when the network suddenly disconnected from the zoom-land, when each of us was in a different time zone, and when... Life happens, but the passion for science and knowledge unites us together.

Let me extend my special thanks to the two amazing undergraduate students, Junyi

Cheng and Xinyuan Lin, whom I have the great pleasure to mentor since 2019. Their passion for physics and curiosity for knowledge have always reminded me of the very reason why I chose to do my Ph.D. in the first place and have again and again rekindled and reinvigorated me during the hard times when I feel lost.

I would also like to thank all the amazing physics professors and teachers I have met along the way. I would not be enticed into the world of science were it not for the curiosity-sparkling, half sci-fi speculations made by my high school teachers, and all their humorous stories and good jokes which turned the learning journey into a cheerful breeze. My particular thanks go to James Loach who opened my door to real-world physics research and to western philosophy in general; to Haiguang Xu with whom I conducted my first baby research project in astrophysics; to Steffen Bass who generously took me in at a very late stage as a rising senior, gave me my first computational project for the study of quark-gluon plasma and heavy-ion collision, and has offered me so much guidance and support ever since then; to Andrew Larkoski, Benjamin Nachman, Mariangela Lisanti, Jesse Thaler, Isabel Garcia Garcia, Ken Van Tilburg, Jessica Howard, Zhengkang Zhang, and many others for their invaluable help and professional advices; and to the numerous professors at Shanghai Jiao Tong University, at Duke, and at UCSB, who have not only taught me how to think like a physicist but also cared a great deal for me as an aspiring female scientist.

Last but certainly not least, my deepest appreciation goes to my family—my parents and grandparents, whose unwavering support and unconditional love have been the very foundation of my strength. Thank you for constantly reminding me how not to overstress and overwork myself, how to set realistic goals and march calmly towards them, how to enjoy the ups and downs and appreciate the little joys in life and stay grounded, and most importantly, how to take a breath and simply *live*. Thank you, especially to my mom, for always being there with me in video calls, even though we are thousands of miles and tens of hours apart!

And finally, a “woof-woof” thanks to my cute little Brownie, who burst into my life like a bright sunray amid the gloomy lockdown. You show me what *happiness* truly is—how simple, how easy, just a loving cuddle and an extra bite of treat!

Curriculum Vitæ
Tianji Cai

Education

2023	Ph.D. in Physics (Expected), University of California, Santa Barbara.
2020	M.A. in Physics, University of California, Santa Barbara.
2017	B.S. in Physics and Mathematics, High Distinction in Physics, Duke University.
2017	B.S. in Physics, Shanghai Jiao Tong University.

Publications

I. Banta, T. Cai, N. Craig, and Z. Zhang, Structures of Neural Network Effective Theories, *preprint*, `arXiv:2305.02334`.

T. Cai, J. Cheng, K. Craig, and N. Craig, Which Metric on the Space of Collider Events?, *Phys. Rev. D 105, 076003*, `arXiv:2111.03670`.

H. Al Ali et al., The Muon Smasher's Guide, *Rep. Prog. Phys. 85 084201*, `arXiv:2103.14043`.

T. Cai, J. Cheng, B. Schmitzer, M. Thorpe, The Linearized Hellinger-Kantorovich Distance, *SIAM Journal on Imaging Sciences Vol. 15, No. 1, pp. 45-83*, `arXiv:2102.08807`.

T. Cai, J. Cheng, K. Craig, and N. Craig, Linearized Optimal Transport for Collider Events, *Phys. Rev. D 102, 116019*, `arXiv:2008.08604`.

Abstract

Optimal Transport for High Energy Physics

by

Tianji Cai

High energy physics, like many other scientific disciplines, has entered an exciting new era of big data, where both particle accelerators at the *energy frontier* and astrophysical surveys at the *cosmic frontier* are producing an enormous amount of data which may hold the very key to the most fundamental questions about nature. Mining such gold inevitably calls for revolutionary designs of ever more powerful and efficient statistical analysis frameworks, while at the same time scientific rigorousness places an additional requirement on the interpretability of any novel model proposed. Among a plethora of available modern machine learning techniques, the theory of *optimal transport* stands out as a distinct approach that is both high performing and mathematically well grounded. By equipping the space of data represented as distributions with a suitable metric, optimal transport replaces *ad hoc* notions of similarity with a well-defined distance, opening up a range of new applications with profound theoretical implications.

This thesis introduces the theory of optimal transport with an eye towards its usage in physics. Special emphasis is put on two particular optimal transport distances which enjoy unique geometric properties. Utilizing their geometric structure, we develop a computationally efficient linearization framework for the two distances and highlight their approximations for discrete distributions encountered in practice. We then showcase the power of this linearized optimal transport framework by applying it to two use cases—one in collider physics at the energy frontier and the other in dark matter astrophysics at the cosmic frontier. As the adoption of optimal transport in high energy physics is still in its

early stage, the present thesis invites the readers to think of other potential applications for their own research.

Permissions and Attributions

1. The content of Chapter 2.1, 2.3, 2.4 and Chapter 3.2, 3.3 is the result of collaboration with Junyi Cheng, Katy Craig, and Nathaniel Craig. This work has previously appeared in the Physical Review D (Phys. Rev. D **102** (2020) 116019).

2. The content of Chapter 2 and Chapter 3.4 is based on work with Junyi Cheng, Bernhard Schmitzer, and Matthew Thorpe, published in SIAM Journal on Imaging Sciences in Vol. 15, No. 1 (2022), pp. 45-83, by the Society for Industrial and Applied Mathematics (SIAM).

3. The content of Chapter 2.5 and Chapter 3.5, 3.6 is the result of collaboration with Junyi Cheng, Katy Craig, and Nathaniel Craig, and previously appeared in the Physical Review D (Phys. Rev. D **105** (2022) 076003).

4. The content of Chapter 3.7 is the result of several ongoing collaborations with Xinyuan Lin, Jessica Howard, Katy Craig, and Nathaniel Craig.

5. The content of Chapter 4 is the result of collaboration with Xinyuan Lin, Nathaniel Craig, as well as early contributions from Mariangela Lisanti and Dylan Folsom.

Contents

Chapter 1

Introduction

Recent advances in artificial intelligence (AI) have revolutionized the way we approach fundamental sciences. In high energy physics (HEP), AI has been applied to a wide range of problems, such as hardware design, lattice quantum chromodynamics (QCD) calculations, collider phenomenology, astrophysics, cosmology, and even string theory [1]. In addition to more traditional machine learning (ML) methods such as boosted decision trees that have been in use in HEP for decades [2, 3], a variety of newly designed deep neural network (NN) architectures are increasingly proving their immense potential to help tackle previously intractable problems in today's big-data era.

Yet approaches employing AI methods, especially neural networks, oftentimes suffer from two main drawbacks: they are hard to interpret and expensive to train. From a theoretical point of view, the former interpretability issue is more concerning, as in science we cannot simply be satisfied with better performances but are obliged to comprehend the underlying reasons. This imperative has triggered another surge of research interests in finding the so-called middle path, that is, to design equally powerful machine learning frameworks which are instead physics-inspired, theoretically-grounded, and thus more amenable to human understanding.

Among such middle paths emerges a distinct approach based on the mathematical theory of optimal transport (OT). Despite its very recent introduction to collider physics in 2019 [4], optimal transport has already witnessed many interesting applications beyond the collider context, stretching itself into fields as far as quantum field theory and dark matter astrophysics. A fast-growing number of high energy physicists, both in the theory community and the experimental community, are getting interested in learning and applying this powerful mathematical tool to diverse physics scenarios.

This thesis is written for such a physics audience. The goal is to present a self-contained, mini-review of the key mathematical concepts in optimal transport theory, so that readers with no prior knowledge can (hopefully) start implementing OT for their own problems after reading the present work and the two illustrative applications therein. More importantly, we wish to convey our enthusiasm for this exciting research program, which has just gained its momentum and beginning to show its fruitfulness.

Indeed in the realm of image analysis, optimal transport has long proven its importance [5, 6, 7, 8, 9, 10]. Naturally, an image can be viewed as a discrete distribution composed of thousands of pixels, with the intensity of each pixel indicating the quantity of "stuff" contained at its particular location. Usually, pixels are not just thrown randomly onto an image of some meaning. Instead, the relative locations and the intensities of individual pixels carry critical information about the content of an image.

More broadly, data in many scientific applications can also be represented in the format of distributions, which, just like images, often contain rich substructure that encodes their very essence. The question then becomes how to best extract such substructure information. By defining a way to move "stuff" around to morph one distribution into another with the least amount of effort, optimal transport provides a mathematically well-grounded distance that can reflect the geometry underlying the data distributions. Let us take a closer look at this statement.

1.1 Euclidean Distance *vs.* Optimal Transport Distances

One may very well wonder why it is necessary to introduce optimal transport in the first place. Why can't we just use the good old Euclidean distance to measure the difference between two data distributions? What is missing in such a naïve Euclidean definition?

Let us consider a very simple example. Imagine our two distributions are nothing but two single Dirac masses. That is, each consists of a single dot with unit mass [1], located at two different positions. We denote them as

$$\mathcal{E} = \delta_x, \qquad \tilde{\mathcal{E}} = \delta_{\tilde{x}}. \tag{1.1}$$

Let us further set up a coordinate system and put the two Dirac masses on the coordinate grid with N bins. For instance, we can specify the locations to be

$$x = (2, 6), \qquad \tilde{x} = (5, 4). \tag{1.2}$$

This step is called "binning" and is illustrated in Figure 1.1, where $\mathcal{E}$ is represented by the blue dot and $\tilde{\mathcal{E}}$ by the red one.

To use the standard image-based approach, one should represent the mass at each grid location, i.e., in each bin, by a vector v (or $\tilde{v}$) $\in \mathbb{R}^N$, where again N is the total

[1] Here "mass" loosely refers to the amount of "stuff" at a given location. For example, it can be "probability mass" in a probability distribution. Later, we will define it more rigorously in terms of measures.

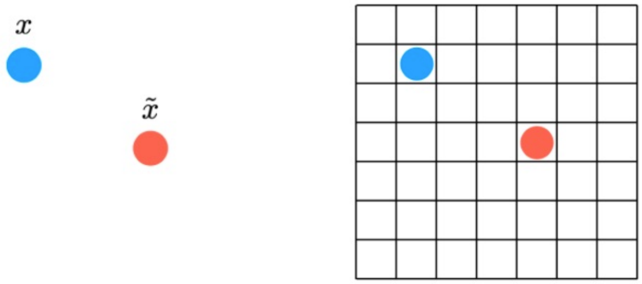


Figure 1.1: Two distributions (blue for $\mathcal{E}$ and red for $\tilde{\mathcal{E}}$) of single Dirac with unit mass. The right subplot places a coordinate grid on the underlying space.

number of bins. In our case, we have

$$\begin{aligned} v &= (0, \cdots, 0, 1_{\text{at bin }(2,6)}, 0, \cdots, 0, 0, \cdots, 0), \\ \tilde{v} &= (0, \cdots, 0, 0, \cdots, 0, 1_{\text{at bin }(5,4)}, 0, \cdots, 0), \end{aligned} \tag{1.3}$$

where the locations of the 1's in the vectors are determined by the locations of the bins where the masses reside. All other entries are simply zero, as there is no mass in those bins.

The image-based distance between the two distributions $\mathcal{E}$ and $\tilde{\mathcal{E}}$ is then given by the Euclidean distance between v and $\tilde{v}$, i.e.,

$$d_{\ell^2(\mathbb{R}^N)}(\mathcal{E}, \tilde{\mathcal{E}}) := \left(\sum_{i=1}^{N} |v_i - \tilde{v}_i|^2 \right)^{1/2} = (1+1)^{1/2} = \sqrt{2}. \tag{1.4}$$

The above equation holds no matter how we move the two Dirac masses around, as long as $x \neq \tilde{x}$. The distance between $\mathcal{E}, \tilde{\mathcal{E}}$ will always be $\sqrt{2}$. In the case where $x = \tilde{x}$, the distance degenerates to 0, as now $v_i = \tilde{v}_i$ for all the ith bins.

Such a distance is obviously not what we want. What we want is something that can at least vary according to the actual locations of the Dirac masses. In other words, we want the new distance defined on the distributions to faithfully reflect the underlying geometry of the space where the distributions themselves live in, which is called the *ground space*. As we have seen, the usual image-based approach is blind to the ground space and fails to provide a satisfying distance.

Now if we think more carefully, there is actually some very familiar concept in physics that we can borrow to define such a distance. If we modify the "force" in the original definition of *work* to be mass, then in our current case the *work-like* distance between $\mathcal{E}$ and $\tilde{\mathcal{E}}$ becomes

$$d_{\text{work-like}}(\mathcal{E}, \tilde{\mathcal{E}}) := m \cdot d_{\ell^2(\mathbb{R}^2)}(x, \tilde{x}) = 1 \cdot \sqrt{(2-5)^2 + (6-4)^2} = \sqrt{13}, \tag{1.5}$$

where $d_{\ell^2(\mathbb{R}^2)}(x, \tilde{x})$ is the distance between $x, \tilde{x}$ in the ground space $\mathbb{R}^2$, denoted as the *ground metric*. Clearly, this work-like distance depends as desired on the location of the Dirac masses in each distribution and therefore is more suitable than the previous image-based approach.

Optimal Transport generalizes the concept of work above to a family of distances that lift the ground metric on the ground space to the set of probability distributions on that space. As we will see later, the ability to preserve spatial information encoded in the ground space is critical to the success of OT distances in a variety of statistical tasks. The classical Euclidean ℓ^2 norm in the image-based approach neglects the geometry of the ground space and is simply not rich enough to handle more sophisticated situations.

1.2 Earth Mover's Distance and Its Modification

Intuitively speaking, optimal transport quantifies the least amount of "work" required to rearrange one distribution to look like the other, be them discrete (such as described above) or continuous. In other words, among all possible rearrangements, it searches for the optimal way to transport one distribution into another; and thus its name. This least effort principle is ubiquitous in physics, sciences, and life in general. Therefore it should not come as a surprise to see optimal transport appearing in widely different problem settings.

In this section, we introduce our very first OT distance, the Earth Mover's Distance (EMD), popularized in the image analysis community [11, 12] and first brought into collider physics [4] to define a distance between collider events. Simply put, the EMD is the minimum cost between two distributions of equal mass, where cost is defined to be the amount of mass moved times the distance by which it is moved.

Let us assume the simple case of two discrete distributions $\mathcal{E} = \sum_i E_i \delta_{x_i}$ and $\tilde{\mathcal{E}} = \sum_j \tilde{E}_j \delta_{\tilde{x}_j}$. That is, each distribution is composed of a bunch of Dirac masses at the given locations. Adopting collider physics language, let's call these Dirac masses *particles*. We further assume that the total masses of the two distributions are the same, i.e., $\sum_i E_i = \sum_j \tilde{E}_j$, in which case we say the distributions are *normalized* and the corresponding optimal transport problem is called *balanced*.

The Earth Mover's Distance is then given by

$$\begin{aligned} \mathrm{EMD}(\mathcal{E}, \tilde{\mathcal{E}}) &:= \min_{\gamma_{ij} \in \Gamma^{\mathrm{EMD}}_{(\mathcal{E}, \tilde{\mathcal{E}})}} \sum_{ij} d_{ij} \gamma_{ij} \qquad (1.6) \\ \Gamma^{\mathrm{EMD}}_{(\mathcal{E}, \tilde{\mathcal{E}})} &:= \left\{ \gamma_{ij} : \gamma_{ij} \geq 0, \ \sum_j \gamma_{ij} = E_i, \ \sum_i \gamma_{ij} = \tilde{E}_j \right\}. \end{aligned}$$

Here d_{ij} is the pre-specified ground metric between particle i in $\mathcal{E}$ and particle j in $\tilde{\mathcal{E}}$. In other words, the user needs to input these ground distances herself, usually in the form of a cost matrix. Then γ_{ij} represents how much mass is moved from particle i to particle j and $\sum_{ij} d_{ij}\gamma_{ij}$ gives the total required effort to transport the distribution $\mathcal{E}$ to $\tilde{\mathcal{E}}$. Notice the three conditions that γ_{ij} needs to satisfy: the mass moved must be nonnegative; the mass moved out of a certain location in $\mathcal{E}$ must be equal to the amount it originally contains; and the mass moved into a given location in $\tilde{\mathcal{E}}$ must also be the same as the amount it originally has. Later, we will rephrase these conditions using the standard optimal transport terminology.

The EMD considers all such rearrangements and outputs the least amount of cost. The corresponding optimal rearrangements can also be obtained, though in general there may be more than one that output the minimum distance. To give a more intuitive illustration of the EMD, we generate two arbitrary discrete distributions on the ground space $\Omega = [-1, 1] \times [-1, 1] \in \mathbb{R}^2$; see the orange and blue dots in Figure 1.2. More general than the previous Figure 1.1, here the mass of every individual dot is usually different from each other and the total masses of the two distributions are set to be unequal. Specifically, the orange distribution has total mass $\sum E_i = 1.8$, whereas the total mass of the blue distribution is $\sum \tilde{E}_j = 1$.

Now in order use the EMD as defined in Equation (1.6), we need to first normalize the two distributions and only afterwards can we calculate the Earth Mover's Distance. The resulting EMD and the corresponding optimal mass rearrangement is presented in the left plot in Figure 1.2. Here the gray lines indicate how much mass is moved from one location to the other with its darkness proportional to the amount of mass moved.

In order not to throw away the potential information encoded in the total mass discrepancy, a simple way is to add to the standard EMD a term that penalizes the corresponding difference. For example, one can define a *modified* Earth Mover's Distance

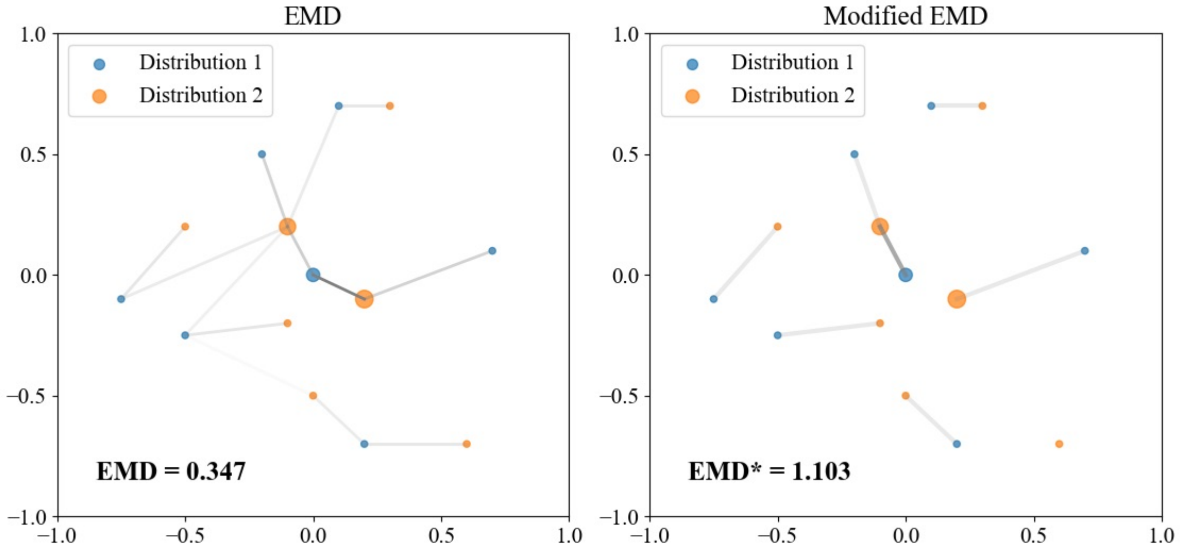


Figure 1.2: Two randomly generated discrete distributions (orange and blue) on the ground space $\Omega = [-1, 1] \times [-1, 1] \in \mathbb{R}^2$, with the size of the dots indicating how much mass there is at a certain location. The total mass of the blue distribution is 1, whereas that for the orange distribution is 1.8. *Left*: The Earth Mover's Distance and the corresponding optimal rearrangement plan between the two distributions as defined in Equation (1.6). *Right:* The modified Earth Mover's Distance and the corresponding optimal rearrangement plan between the two distributions as defined in Equation (1.7).

as

$$\mathrm{EMD}^*(\mathcal{E},\tilde{\mathcal{E}}) := \min_{\gamma_{ij}\in\Gamma^{\mathrm{EMD}^*}_{(\mathcal{E},\tilde{\mathcal{E}})}} \sum_{ij} d_{ij}\gamma_{ij} + \alpha\left|\sum_i E_i - \sum_j \tilde{E}_j\right| \tag{1.7}$$

$$\Gamma^{\mathrm{EMD}^*}_{(\mathcal{E},\tilde{\mathcal{E}})} := \left\{\gamma_{ij} : \gamma_{ij} \geq 0,\ \sum_j \gamma_{ij} \leq E_i,\ \sum_i \gamma_{ij} \leq \tilde{E}_j, \sum_{ij}\gamma_{ij} = E_{\min}\right\},$$

where $E_{\min} := \min\left(\sum_i E_i, \sum_j \tilde{E}_j\right)$ and α is a free parameter that controls the relative importance between the standard EMD term and the total mass difference term. The modified EMD, denoted as EMD*, is sometimes also called Earth Mover's Distance, or else Energy Mover's Distance in the literature. Here we simply put an adjective in front to avoid any potential confusion.

Notice that the legitimate rearrangements in this case satisfy a different set of constraints than in the standard EMD case in Equation (1.6). This will be explained in more detail later. The resulting optimal rearrangement and the corresponding distance for EMD* is illustrated in the right plot in Figure 1.2, where α is set to 1 for simplicity. As can be seen, the rearrangement plans are rather different for EMD and EMD*. Additionally, the difference between the two distances is around 0.76, which is about the same as the total mass difference of 0.8, as should be the case.

1.3 Related Works

The above EMD and its modified version were first introduced to collider physics in [4], as an attempt to provide the space of collider events with a suitable metric. It is then further developed in [13], where the simple geometric language of EMD is shown to elegantly unify diverse concepts in jet physics and quantum field theory. Among other things, EMD allows for a novel definition of infrared and collinear safety; provides a new perspective on a number of existing collider observables as the minimum distance to a certain manifold; implies inequalities satisfied nonperturbatively by jet variables; casts common jet definitions as a problem of finding the closest N-particle approximation; gives new expressions to various pileup mitigation strategies; and may even enable a precise definition of a distance between theories. The two groundbreaking works lay the foundation for the adoption of optimal transport based techniques in collider physics.

Since then, the EMD and its many variants have flourished along multiple directions. It has been applied to distance-based analysis of jets in CMS Open Data [14], to the definition of a new “event isotropy” shape variable [15], (with suitable generalization) to discrimination at the full event level [16], to data embedding into lower-dimensional spaces with simple metrics [17], to the calibration of stochastic simulations [18], to specific new physics searches such as CP violation [19], to the generalization of the notion of event and jet shape observables and their efficient computation [20], and to the introduction of continuous jet grooming for noise removal [21]. Recent work [22] has also modified the input ground space of the EMD to be one-dimensional spectral functions with basic symmetries of collider data built-in and has accordingly defined a novel 1D *spectral EMD*, which enjoys enhanced analytical calculability from first-principles.

Another promising avenue being pursued is to use the EMD—or in general, other optimal transport distances—as a more sophisticated and physics-inspired loss function of various neural network architectures. For example, such OT-based metrics are employed in autoencoders for anomalous jet tagging [23, 24], for the construction of a metrized latent space of collider events [25], and for fast simulators that directly link experimental data with the underlying theoretical models [26]. It has also been used as the cost function of a self-supervised graph neural network with attention mechanisms to identify particles of the primary collision event from large pileup contaminiation [27]. A number of other metrics for collider events have also been explored in [28]. Broadly speaking, the many applications of the EMD highlight the potential relevance of tools from the theory of optimal transport for collider physics.

Beyond collider phenomenology, optimal transport has also found surprising and profound connections with various formal studies of quantum field theories, gravity, and string theory. In specific, [29] has reformulated exact renormalization group flow in the language of optimal transport gradient flow, providing a novel viewpoint for further works on an information theoretic approach to renormalization [30, 31]. In [32], it has been demonstrated using optimal transport that the Einstein equations are equivalent to a simple concavity property of an entropy in the Wasserstein space. An extension for gravity compactification has also been provided, where a newly developed framework of optimal transport in space with negative dimensions is considered. For the latter, the tools of OT subsequently prove useful in deriving general bounds on the masses of the Kaluza-Klein particles [33].

1.4 Thesis Outline

Hopefully, the above discussions have offered an intuitive idea of what optimal transport is, as well as some motivations for its introduction to high energy physics. The rest of the thesis is then devoted to the discussions of two specific HEP fields where optimal transport has proven or is expected to show great potential. But first, we need to better understand the theory of optimal transport itself and develop the necessary analytical tools customized to our later physics applications.

This is what Chapter 2 is all about. The focus here is always to maintain a clear flow of the narrative for ease of understanding, and therefore intuitive arguments are preferred over mathematically rigorous proofs. Whenever possible, we refer the readers to the original articles for complete expositions. As the field of optimal transport is rich with many active research happening at the time of writing, it is not possible to do justice to such a vast mathematical field in one short chapter. Fortunately, many good review articles and textbooks exist, which will be pointed out in due time. While a certain part of it deals with standard textbook materials, the main body of Chapter 2 consists of new results mostly derived from our own published papers, especially [34].

We then begin our voyage into the two physics applications of optimal transport. Chapter 3 focuses on collider physics, where optimal transport is used to define a physically meaningful metric on the space of collider events. This chapter combines the results from our three publications [34, 35, 36] and additionally discusses our several current endeavors to upgrade the methodological framework of optimal transport for collider physics applications.

Chapter 4 marches into the still uncharted territory of applying optimal transport to dark matter astroparticle physics, in specific, the study of dark matter halos. Although the physics use case is on a more speculative front, the statistical framework itself is full-fledged and awaits diverse applications. We will explain our novel framework in full detail and present the preliminary results. A discussion of problems observed in our study is also included, paving the way for further developments which will hopefully resolve these issues. Notice that none of the results in Chapter 4 has yet made its way to publication and everything is still under active research at the time of writing. Therefore, major revisions to the presented results are possible and we invite the readers to also think about other potential usages of the framework developed here.

Finally, Chapter 5 concludes the thesis by giving some final thoughts on the future of this promising research direction. As a young, burgeoning field full of promises, we look forward to seeing many more studies of optimal transport for high energy physics in the near future, both from the perspective of offering a fresh theoretical understanding and from the practical side of upgrading our data analysis toolbox. Hopefully, this present thesis can be of some use to the interested readers by offering an accessible mathematical treatment, by presenting two compelling physics applications, and by pointing out a wealth of valuable references for further reading.

Chapter 2

Optimal Transport Theory in a Nutshell

First proposed by French mathematician Gaspard Monge (1746–1818), optimal transport theory (OT) emerged as an engineering problem where the concern was about how to move a pile of sand to another of the same volume with the least amount of work. This defines the fundamental question of optimal transport, which, despite its apparent simplicity, is surprisingly hard to formulate—let alone solve—in a mathematically rigorous way. It wasn't until the mid 20th century that a group of mathematicians revisited the problem and firmly rooted it in optimization theory. Among them was the Soviet mathematician Leonid Kantorovich, who rephrased OT in the language of general measure theory.

Simply put, OT associates a "global" cost to each possible way of morphing, or *transporting*, one distribution into the second by considering how much it costs to "locally" move an infinitesimal amount from one location to another. In addition to defining a proper distance between distributions, the smallest cost also gives rise to rich geometric structures on the space of distributions by inheriting the key properties of the underlying ground space—this advantage of OT is already highlighted in the *Introduction*.

Recent years have again witnessed a surge of interest in optimal transport, thanks to its connection to and usefulness in many statistical machine learning problems. Contributing to this rapid spread of OT is the emergence of approximate OT solvers which can be more easily upscaled to large problem dimensions. At the time of writing, the field of optimal transport remains highly active—new extensions to the framework are being devised, which further empower this incredible mathematical tool on the practical front.

In Chapter 1, we have had our first encounter with optimal transport via the introduction of the Earth Mover's Distance (EMD) and its modification. Properly speaking, EMD is only one example of a family of *balanced* optimal transport distances defined between equally normalized distributions, known as the *p-Wasserstein distances.* The p-Wasserstein may be modified to accommodate distributions with different total masses. One way is to incorporate an additional term to account for differences in the total mass, as adopted in the modified EMD definition.

Yet this extension is far from unique. There are many possible approaches to the *unbalanced* optimal transport problem. Of course, at the end of the day, the "best" OT distance depends on many relevant criteria for a given application. Practical considerations include simplicity, robustness, and computational speed, while theoretical considerations favor geometric interpretability.

Despite the richness of the subject, in this chapter we simply group optimal transport into two broad categories, balanced OT and unbalanced OT. Section 2.1 focuses on a special balanced optimal transport distance, the 2-Wasserstein (W_2) distance, whereas Section 2.2 examines its unbalanced counterpart, the Hellinger-Kantorovich (HK) distance, whose structure is more complex yet, in a way, parallel to that of the W_2 distance.

Then in Section 2.3, we discuss the practical challenge of calculating optimal transport metrics. Despite the acceleration provided by many approximate solvers, the high computational cost of OT still poses a serious limitation to its real-world adoption. Fortunately, both the W_2 and the HK distances enjoy a special geometrical structure that lends themselves easily to linearization. The following two sections 2.4 and 2.5 expound the idea of linearization for the two particular distances and develop the essential framework for their later applications to discrete distributions, especially under the collider physics context in Chapter 3. Finally, Section 2.6 presents a simple example to illustrate the above mathematical concepts from a practical point of view. It serves as a sandbox to test the linearized optimal transport framework before trying it out on more complicated datasets.

The present chapter lays the mathematical groundwork necessary for the understanding and application of optimal transport for the rest of the thesis. Mainly based on the author's three publications [35, 34, 36], the current presentation avoids detailed proofs and makes certain modifications to our previous notations. As any rich and well-developed theory, optimal transport certainly is in no shortage of excellent learning resources. We refer the reader to the textbooks by Peyré and Cuturi [37] (the author's primary source of reference), Ambrogio, Gigli, and Savaré [38], Santambrogio [39], and Villani [40, 41] for further background materials and detailed expositions. Also see [42] for an overview of the recent application of optimal transport in image and signal analysis, and [43] for a survey of optimal transport applications in machine learning.

Before we start, let us first briefly outline some preliminaries for optimal transport theory and establish the key notations used throughout this chapter and beyond.

2.0.1 Preliminaries and Notations

Since optimal transport distances are defined on distributions, we first need a mathematically rigorous way to describe distributions themselves. This is given by measure theory. Roughly speaking, all the distributions we have been talking about can be seen as measures on $\mathbb{R}^d$. For example, a 1D distribution with unit total mass is called a probability measure on $\mathbb{R}^1$. Here we review some basic concepts and notations in measure theory relevant to the later setup of optimal transport. For a complete treatment on measure theory, please refer to any standard mathematical textbook.

Let Ω denote a convex, closed, bounded subset of $\mathbb{R}^d$ with non-empty interior. For a compact metric space X, $\mathrm{C}(X)$ denotes the space of continuous real-valued functions over X, and similarly $\mathrm{C}(X)^n$ denotes the space of continuous $\mathbb{R}^n$-valued functions. In most applications, X is the metric space $\mathbb{R}^d$ itself (or the metric space of a subset of $\mathbb{R}^d$). We use $\mathcal{M}(X)$ for the space of signed Radon measures over X, $\mathcal{M}_+(X)$ for the space of non-negative Radon measures, $\mathcal{M}_1(X)$ for the set of probability measures (also denoted as $\mathcal{P}(X)$), and $\mathcal{M}(X)^n$ for the space of $\mathbb{R}^n$-valued measures. We identify $\mathcal{M}(X)$ and $\mathcal{M}(X)^n$ with the dual spaces of $\mathrm{C}(X)$ and $\mathrm{C}(X)^n$.

The Lebesgue measure on various domains is denoted by $\mathcal{L}$, where a subscript is added when the domain is not clear from the context. The set of Lebesgue-absolutely continuous measures is denoted by $\mathcal{M}_{\mathcal{L}}(X)$, and we further define $\mathcal{M}_{1,\mathcal{L}}(X) := \mathcal{M}_{\mathcal{L}}(X) \cap \mathcal{M}_1(X)$.

In general, measures can be continuous or discrete and optimal transport theory is flexible enough to deal with both types of measures simultaneously within the same framework. As numerical applications almost always deal with discrete measures only, we treat the special case of discrete measures separately and in much more detail. In this case, we use upper case curly English letters, for instance, $\mathcal{E}$ and $\tilde{\mathcal{E}}$, alluding to the later collider events they stand for. Such discrete measures are composed of a set of Diracs, each with a certain amount of mass located at a given position in the underlying space X. We therefore write $\mathcal{E} = \sum_i E_i \delta_{x_i}$ and $\tilde{\mathcal{E}} = \sum_j \tilde{E}_j \delta_{\tilde{x}_j}$, where the $E_i, \tilde{E}_j$ are the masses of the individual Diracs located at $x_i, \tilde{x}_j$, respectively. Again borrowing collider physics language, we oftentimes call these individual Diracs as *particles* in a given event. In the case where $\sum_i E_i = 1$, we say the event (measure) $\mathcal{E}$ is normalized. Finally, the letter $\mathcal{R}$ is usually reserved for a discrete reference measure, i.e., $\mathcal{R} = \sum_i R_i \delta_{y_i}$.

If instead we are concerned about general measures (discrete and continuous), we resort to lower case greek letters such as μ, ν. We usually add a subscript for different measures and often reserve μ_0 for a continuous reference measure. Now given two probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, a measurable function $\mathbf{t} : \mathbb{R}^d \to \mathbb{R}^d$ *transports μ onto ν* if $\nu(B) = \mu(\mathbf{t}^{-1}(B))$ for all measureable sets $B \subseteq \mathbb{R}^d$. We call ν the *push-forward of μ under* $\mathbf{t}$ and write $\nu = \mathbf{t}_{\#}\mu$. For historical reasons, it is conventional in the field of optimal transport to think of the amount of measure μ gives to a measurable set B as the *mass of B with respect to μ* and to interpret a measurable function $\mathbf{t}$ as a *transport map* that *rearranges the mass in μ to look like ν.*

Given a probability measure on a product space, for example $\boldsymbol{\gamma} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$, its *marginals* are given by the pushforward of the measure through the projections on each component of the product. For example, if $\pi^2 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is the projection onto the second component of $\mathbb{R}^d \times \mathbb{R}^d$, then $\pi^2_{\#}\boldsymbol{\gamma}$ is the *second marginal* of γ.

As always, the Euclidean norm and inner product on $\mathbb{R}^d$ are denoted by $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ respectively. We will also use $\|\cdot\|$ as the norm on $\mathbb{C}$ and the total variation norm on measures. For positive measures $\mu \in \mathcal{M}_+(\Omega)$, we just have $\|\mu\| = \mu(\Omega)$.

2.1 Balanced OT: The 2-Wasserstein Distance

First we consider the case of *balanced* optimal transport, where the total masses of the two distributions under comparison are the same. Without loss of generality, one can always normalize each distribution so that its total mass becomes one.

Below we first introduce the p-Wasserstein metrics in the simpler setting of discrete measures most suitable for the later physics applications in Chapter 3 and Chapter 4. We then present two equivalent formulations of the special 2-Wasserstein (W_2) distance for general measures in more mathematical detail. An example calculation of W_2 between Dirac measures is given, after which the geodesics for W_2 is generalized.

2.1.1 p-Wasserstein Distance for Discrete Measures

Assume we have two discrete measures $\mathcal{E} = \sum_i E_i \delta_{x_i}$ and $\tilde{\mathcal{E}} = \sum_j \tilde{E}_j \delta_{\tilde{x}_j}$, possibly with different total number of particles. Without loss of generality, one may assume that both measures are *normalized*, i.e., $\sum_i E_i = \sum_j \tilde{E}_j = 1$, and we always have $E_i, \tilde{E}_j \geq 0$. In the context of collider physics, for example, $\mathcal{E}$ may represent an event or a jet consisting of n particles at locations x_i in a rectangular domain Ω on the unfolded detector plane. [1] The mass E_i then denotes the energy (or the transverse momentum in the case of a hadron collider) of the ith particle in the event.

Given two such measures, the p-Wasserstein distance with $p \geq 1$ between them is

[1] While the detector on which the collision data is recorded is a cylinder, due to the fact that we will later translate jets clustered with unit radius parameter to be centered at the origin, we may neglect the periodic boundary conditions in the azimuthal angle and consider the underlying domain to be a rectangle.

defined by

$$\mathrm{W}_p(\mathcal{E},\tilde{\mathcal{E}}) = \min_{g_{ij}\in\Gamma(\mathcal{E},\tilde{\mathcal{E}})} \left(\sum_{ij} g_{ij}\|x_i - \tilde{x}_j\|^p\right)^{1/p}, \tag{2.1}$$

$$\Gamma(\mathcal{E},\tilde{\mathcal{E}}) = \left\{ g_{ij} : g_{ij} \geq 0,\ \sum_j g_{ij} = E_i,\ \sum_i g_{ij} = \tilde{E}_j \right\},$$

where $\|x_i - \tilde{x}_j\|$ denotes the distance on the underlying space Ω, which will be refered to as the *ground metric* on the *ground space.* When $p = 1$, the above definition reduces to the Earth Mover's Distance. When $p = 2$, we have the special case of the 2-Wasserstein (W_2) distance, also known as the Monge-Kantorovich distance.

One interpretation of the p-Wasserstein distance is that it represents the minimal amount of "effort" required to rearrange the distribution of mass in $\mathcal{E}$ to match $\tilde{\mathcal{E}}$. In this case, g_{ij} represents the amount of mass moved from particle i in $\mathcal{E}$ to particle j in $\tilde{\mathcal{E}}$, and $\|x_i - \tilde{x}_j\|^p$ represents the "cost" of moving mass between the two locations in the ground space.

With this interpretation, $\Gamma(\mathcal{E},\tilde{\mathcal{E}})$ is the set of possible ways to rearrange $\mathcal{E}$ to look like $\tilde{\mathcal{E}}$, known as the set of *transport plans.* The viable transport plans must satisfy the following conditions: any rearrangement g_{ij} can only move nonnegative amounts of mass; the total amount of mass moved from a fixed particle i in $\mathcal{E}$ to all of the particles in $\tilde{\mathcal{E}}$ must coincide with the original mass E_i; and, symmetrically, the total amount of mass moved from all of the particles in $\mathcal{E}$ to any fixed particle j in $\tilde{\mathcal{E}}$ must coincide with $\tilde{E}_j$. In other words, $\Gamma(\mathcal{E},\tilde{\mathcal{E}})$ must marginalize to $\mathcal{E}$ and $\tilde{\mathcal{E}}$, respectively.

This gives the "original" static Kantorovich formulation of W_2 for the case of discrete measures.

2.1.2 W_2 Distance: Kantorovich Formulation

More generally, for continuous probability measures $\mu_0, \mu_1 \in \mathcal{M}_1(\Omega)$, the 2-Wasserstein distance is given by

$$\mathrm{W}_2(\mu_0,\mu_1)^2 = \inf \left\{ \int_{\Omega^2} \|x_0 - x_1\|^2 \,\mathrm{d}\pi(x_0,x_1) + \sum_{i\in\{0,1\}} \iota_{\{\mu_i\}}(P_{i\sharp}\pi) \,\middle|\, \pi \in \mathcal{M}_1(\Omega^2) \right\} \tag{2.2}$$

where $\iota_{\{\mu_i\}}$ denotes the indicator function of $\{\mu_i\}$, i.e.,

$$\iota_{\{\mu_i\}}(\nu) := \begin{cases} 0 & \text{if } \nu = \mu_i, \\ +\infty & \text{else.} \end{cases}$$

One can check that it returns the same definition as Equation (2.1) in the case of discrete measures. Here $P_i : \Omega \times \Omega \to \Omega$, $(x_0, x_1) \mapsto x_i$ are the projections from the product space onto the marginals. The set $\Pi(\mu_0,\mu_1) = \{\pi \in \mathcal{M}_1(\Omega^2) \,|\, P_{i\sharp}\pi = \mu_i \text{ for } i = 0,1\}$ is the set of transport plans or couplings between μ_0 and μ_1, which plays the same role as $\Gamma(\mathcal{E},\tilde{\mathcal{E}})$ in the discrete case with g_{ij} now replaced by $\mathrm{d}\pi(x_0,x_1)$ for infinitesimal amount of mass.

It is well known that W_2 is a metric on $\mathcal{M}_1(\Omega)$ and minimal π in Equation (2.2) exist but is not unique in general. We denote the set of all $\pi \in \mathcal{M}_1(\Omega^2)$ that minimize Equation (2.2) for $\mathrm{W}_2(\mu_0,\mu_1)$ by $\Pi_{\mathrm{opt}}(\mu_0,\mu_1)$ called the *optimal transport plans.* Furthermore, we say that a plan $\pi \in \Pi(\mu_0,\mu_1)$ is *induced by a transport map* if there exists a measurable function $\mathbf{t} : \Omega \to \Omega$ so that $\pi = (\mathbf{id} \times \mathbf{t})_\sharp \mu_0$, where $\mathbf{id}(x) = x$ is the identity mapping.

We devote particular attention to the case that one of the measure, say μ_0, is Lebesgue-absolutely continuous, i.e., $\mu_0 \ll \mathcal{L}$. As $\Omega \in \mathbb{R}$, this means that μ_0 does not give mass to sets of $(d-1)$-dimensional Hausdorff measure; in other words, the measure does not concentrate on small sets. In this case, the minimizer π becomes unique for any μ_1 and is induced by a transport map [44]. This transport map is unique (up to sets of μ_1 measure zero), and we refer to it as the *optimal transport map* from μ_0 to μ_1, denoted $\mathbf{t}_{\mu_0}^{\mu_1}$ [45].

The function $x \mapsto \mathbf{t}_{\mu_0}^{\mu_1}(x)$ represents where mass starting at location x in the source measure μ_0 is sent in the target measure μ_1, in order to rearrange the mass from μ_0 into μ_1 with the least amount of effort. Note that a necessary condition for such an optimal transport map to exist is that an optimal rearrangement of μ_0 to μ_1 does not *split mass*; that is, all mass starting at a specific location in μ_0 must be sent to the same location in μ_1.

2.1.3 W$_2$ Distance: Benamou–Brenier Formulation

There is another dynamic formulation of the 2-Wasserstein distance by Benamou and Brenier [46]. Equivalence between the static and the dynamic formulations for W$_2$ is a classical result in optimal transport theory. Proofs can be found, for instance, in [40, Theorem 8.1] and [39, Theorem 5.28], whereas existence, uniqueness and Monge-map structure of minimizers are also treated in [40, 39, 47].

In the dynamic formulation, W$_2$ is computed by minimizing an action functional over solutions to the continuity equation. Let $\mathcal{CE}(\mu_0, \mu_1)$ denote the set of solutions for the continuity equation on $[0, 1] \times \Omega$. That is, $\mathcal{CE}(\mu_0, \mu_1)$ contains the pairs of measures $(\rho, \omega) \in \mathcal{M}([0, 1] \times \Omega)^{1+d}$, where ρ interpolates between μ_0 and μ_1 and that solve

$$\partial_t \rho + \nabla \omega = 0 \tag{2.3}$$

in a distributional sense. More precisely, we require for all $\phi \in C^1([0,1] \times \Omega)$ that

$$\int_{[0,1]\times\Omega} \partial_t \phi \,\mathrm{d}\rho + \int_{[0,1]\times\Omega} \nabla\phi \cdot \mathrm{d}\omega = \int_{\Omega} \phi(1,\cdot)\,\mathrm{d}\mu_1 - \int_{\Omega} \phi(0,\cdot)\,\mathrm{d}\mu_0. \tag{2.4}$$

The action functional $J_{\mathrm{W}} : \mathcal{M}([0,1] \times \Omega)^{1+d} \to \mathbb{R} \cup \{\infty\}$ is then given by

$$J_{\mathrm{W}}(\rho,\omega) := \begin{cases} \int_{[0,1]\times\Omega} \|\frac{\mathrm{d}\omega}{\mathrm{d}\rho}\|^2 \mathrm{d}\rho & \text{if } \rho \geq 0, \omega \ll \rho \\ +\infty & \text{else.} \end{cases} \tag{2.5}$$

We can now define the W_2 distance as

$$\mathrm{W}_2(\mu_0,\mu_1)^2 := \inf \left\{ J_{\mathrm{W}}(\rho,\omega) | (\rho,\omega) \in \mathcal{CE}(\mu_0,\mu_1) \right\}, \tag{2.6}$$

where the minimizers (ρ,ω) are referred to as *constant speed geodesics* between μ_0 and μ_1 with respect to W_2.

It is often convenient to describe the measures $(\rho,\ \omega)$ via their disintegration with respect to time, as their time-marginals are Lebesgue-absolutely continuous when $J_{\mathrm{W}} < \infty$. For instance, we can define a $\rho_t \in \mathcal{M}(\Omega)$ with $t \in [0,1]$ via

$$\int_{[0,1]\times\Omega} \phi \,\mathrm{d}\rho := \int_{[0,1]} \int_{\Omega} \phi(t,\cdot)\,\mathrm{d}\rho_t \,\mathrm{d}t \tag{2.7}$$

for all $\phi \in \mathrm{C}([0,1] \times \Omega)$. We thus write $\rho \equiv \rho_t \otimes \mathrm{d}t$. We will proceed for ω and other measures in a similar way.

2.1.4 Geodesics for $\mathbf{W}_2$

Let us illuminate the above formalism with a simple example. We consider two Dirac measures $\mathcal{E} = \delta_{x_0}$ and $\tilde{\mathcal{E}} = \delta_{x_1}$. That is, both consist of a single unit-mass particle located either at x_0 for $\mathcal{E}$ or x_1 for $\tilde{\mathcal{E}}$.

Using the static formulation, it is straightforward to see that the 2-Wasserstein distance between them is

$$\mathrm{W}_2(\mathcal{E}, \tilde{\mathcal{E}})^2 = \mathrm{W}_2(\delta_{x_0}, \delta_{x_1})^2 = \|x_0 - x_1\|^2. \tag{2.8}$$

We simply move the particle from x_0 to x_1. This gives the only and the optimal transport plan from $\mathcal{E}$ to $\tilde{\mathcal{E}}$.

Now let

$$X(x_0, x_1; t) = (1-t)\, x_0 + t\, x_1\,, \tag{2.9}$$

which gives a line between x_0 and x_1 parametrized by a time parameter $t \in [0,1]$ for fixed $x_0, x_1 \in \Omega$.

One can now obtain the unique constant speed geodesic between $\mathcal{E}$ and $\tilde{\mathcal{E}}$ for the W_2 metric, i.e.,

$$\rho_t = \delta_{X(x_0,x_1;t)} \qquad\qquad \omega_t = \delta_{X(x_0,x_1;t)} \cdot \partial_t X(x_0, x_1; t)\,. \tag{2.10}$$

Obvious to see, ρ_t gives the moving measure at time t and ω_t gives its speed which is constant. Therefore, a Dirac-to-Dirac geodesic in W_2 consists of a single Dirac traveling along the constant speed line $X(x_0, x_1; \cdot)$ in Ω. See Figure 2.1 for an illustration.

For general measures μ_0, μ_1, we can compute the constant speed geodesics for W_2 from the transport plans π by a superposition of Dirac-to-Dirac geodesics. The intuition behind this is as follows.

If $\pi(x_0, x_1) > 0$ for a pair $(x_0, x_1) \in \Omega^2$, a certain amount of mass indicated by π is moved from position x_0 to position x_1. Breaking it into infinitesimal parts, each is then

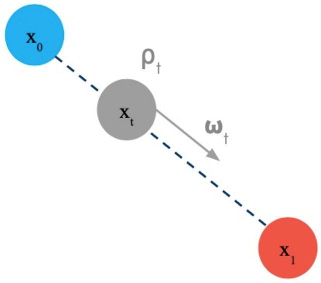


Figure 2.1: An illustration for the Dirac-to-Dirac geodesic in W_2.

the same as the above Dirac case and thus travels along a Dirac-to-Dirac geodesic from x_0 to x_1. Therefore, the geodesic between μ_0 and μ_1 is just the superposition of all these Dirac-to-Dirac geodesics.

A precise formula for (ρ, ω) can be obtained. Let $\mu_0, \mu_1 \in \mathcal{M}_1(\Omega)$ and let $\pi \in \mathcal{M}_+(\Omega \times \Omega)$ be a corresponding minimizer of Equation (2.2), i.e., the optimal couplings. Again define X by Equation (2.9). Then a constant speed geodesic between μ_0 and μ_1 is given by

$$\rho_t := \int_{\Omega^2} \delta_{X(x_0,x_1;t)} \,\mathrm{d}\pi(x_0,x_1) = X(\cdot,\cdot;t)_\sharp \pi \tag{2.11}$$

where $X(\cdot,\cdot;t)_\sharp \pi$ denotes the push-forward of π under X with the t-argument fixed, and

$$\omega_t := \int_{\Omega^2} \left[\delta_{X(x_0,x_1;t)} \cdot \partial_t X(x_0,x_1;t)\right] \,\mathrm{d}\pi(x_0,x_1) = X(\cdot,\cdot;t)_\sharp \left(\partial_t X(\cdot,\cdot;t) \cdot \pi\right). \tag{2.12}$$

It is not hard to see that these formulas are a superposition of formulas for the Dirac measures. For more details on the W_2 geodesics, please refer to [39, Theorem 5.27].

2.2 Unbalanced OT: The Hellinger-Kantorovich Distance

The p-Wasserstein distances are only meaningfully defined for measures of equal mass, that is, the mass distributions must be matched exactly. In addition to the obvious limitation it puts on the measures, this requirement also makes the optimal coupling susceptible to noise in the form of small but non-local mass fluctuations. Consequently, small perturbations can suppress and wash away more relevant features, making the optimal transport distances less effective in reflecting the key features of the underlying data measures.

In the past few years, there has been substantial interest in generalizing the balanced optimal transport metrics to measures with unequal total mass, known as the *unbalanced* optimal transport problem [48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58]. In particular, the recently proposed Hellinger–Kantorovich (HK) distance [52, 59, 53, 55] defines a metric by allowing for the creation and destruction of mass within the OT framework. This HK distance enjoys many similar geometric properties as the W_2 distance and can be formulated in a way parallel to the development of the W_2 distance in the previous section.

Of course, the Hellinger–Kantorovich distance is only a particular variant of the unbalanced transport problem and many other models to combine transport and creation/destruction are conceivable. See, for instance, [60, 58, 49, 61]. Some discussion is also provided in [53].

We begin this section by recalling the notion of *partial optimal transport*, already briefly introduced in Chapter 1. We then extend it to a full class of *unbalanced optimal transport* using the dynamic formulation, with a special focus on the Hellinger–Kantorovich distance.

Thereafter, we dive into the HK distance and present the two equivalent formalisms, this time first the Benamou–Brenier-type formulation and then the Kantorovich-type formulation. We then emphasizes the global mass rescaling behavior of the HK distance, pointing out the importance of local mass discrepancy for unbalanced optimal transport. Finally, we again present an example of Dirac masses and work out the geodesics of HK for general measures.

As the mathematics for HK is much more involved, we do not attempt to give rigorous proofs and only results relevant to later usages are presented. For full mathematical details, please consult [34] and the references therein.

2.2.1 Partial Optimal Transport

The first unbalanced optimal transport metric considered in collider physics is the modified Earth Mover's Distance for discrete measures studied in [4] and explained in Chapter 1. In this case, for fixed $R \geq \max_{ij} d_{ij}/2$, the distance between two discrete measures $\mathcal{E}, \tilde{\mathcal{E}}$ is

$$\mathrm{EMD}^{*,R}(\mathcal{E}, \tilde{\mathcal{E}}) = \min_{\gamma_{ij} \in \Gamma^{\mathrm{EMD}^*}_{\leq(\mathcal{E},\tilde{\mathcal{E}})}} \frac{1}{R} \sum_{ij} d_{ij} \gamma_{ij} + \left| \sum_i E_i - \sum_j \tilde{E}_j \right|, \tag{2.13}$$

where a transport plan γ_{ij} belongs to the set $\Gamma^{\mathrm{EMD}^*}_{\leq(\mathcal{E},\tilde{\mathcal{E}})}$ in case it satisfies the following four criteria:

1. $\gamma_{ij} \geq 0$,
2. $\sum_j \gamma_{ij} \leq E_i$,
3. $\sum_i \gamma_{ij} \leq \tilde{E}_j$, and
4. $\sum_{ij} \gamma_{ij} = \min\left(\sum_i E_i, \sum_j \tilde{E}_j\right)$.

These criteria ensure that (1) the amount of mass moved between any two particles is always nonnegative; (2) the maximum amount of mass that can be moved from location i in $\mathcal{E}$ to any location in $\tilde{\mathcal{E}}$ is E_i; (3) the maximum amount of mass that can be moved to location j in $\tilde{\mathcal{E}}$ from any location in $\mathcal{E}$ is $\tilde{E}_j$; and (4) the total mass that is moved equals the total mass of whichever event has smaller mass. If strict inequality holds in constraint 2, we will say $E_i - \sum_j \gamma_{ij}$ mass has been *destroyed* at x_i, and if strict inequality holds in constraint 3, we will say $\tilde{E}_j - \sum_i \gamma_{ij}$ mass has been *created* at $\tilde{x}_j$.

Note that in the above definition, when the measures $\mathcal{E}$ and $\tilde{\mathcal{E}}$ have equal total mass, $\text{EMD}^{*,R}(\mathcal{E}, \tilde{\mathcal{E}}) = \frac{1}{R} W_1(\mathcal{E}, \tilde{\mathcal{E}})$, i.e., we recover the standard balanced EMD (with an extra factor of $1/R$).

In fact, this modified EMD is a special case of the *partial transport distance* studied by Georgiou, Karlsson, and Takyar [62], Caffarelli and McCann [58], Figalli [48], and Piccoli and Rossi [49, 50]. For $\kappa > 0$, $p \geq 1$, define

$$\begin{aligned} &T_p^\kappa(\mathcal{E}, \tilde{\mathcal{E}}) \\ &= \min_{\gamma_{ij} \in \Gamma_{\leq(\mathcal{E}, \tilde{\mathcal{E}})}} \left(\Sigma_{ij} d_{ij}^p \gamma_{ij} \right)^{1/p} + \frac{\kappa}{2} \left(|\Sigma_i E_i - \Sigma_{ij} \gamma_{ij}| + \left| \Sigma_j \tilde{E}_j - \Sigma_{ij} \gamma_{ij} \right| \right), \end{aligned} \tag{2.14}$$

where a transport plan γ_{ij} belongs to the set $\Gamma_{\leq(\mathcal{E}, \tilde{\mathcal{E}})}$ in case it satisfies criteria (1, 2, 3) above.

The two main differences between $\text{EMD}^{*,R}$ and T_p^κ are that, first, the partial transport distances allow $p \geq 1$ and, second, they permit the amount of mass that is rearranged from $\mathcal{E}$ to $\tilde{\mathcal{E}}$ to differ from the total mass of whichever event has smaller mass. To see this, assume without loss of generality that $\mathcal{E}$ has smaller total mass, $\sum_i E_i \leq \sum_j \tilde{E}_j$. The distance $\text{EMD}^{*,R}$ requires that all of the mass in $\mathcal{E}$ be rearranged: exactly $\sum_j \tilde{E}_j - \sum_i E_i$ mass is created in $\tilde{\mathcal{E}}$, and no mass is destroyed. On the other hand, T_p^κ allows for $\sum_{ij} \gamma_{ij} \in (0, \min(\sum_i E_i, \sum_j \tilde{E}_j))$ mass to be rearranged: $\sum_i E_i - \sum_{ij} \gamma_{ij}$ mass is destroyed in $\mathcal{E}$, and $\sum_j \tilde{E}_j - \sum_{ij} \gamma_{ij}$ mass is created in $\tilde{\mathcal{E}}$.

Next we show why, for $\kappa = 2R \geq \max_{ij} d_{ij}$, $\text{EMD}^{*,R}$ coincides with (a constant multiple of) T_1^κ. First, note that the EMD^* constraint set is a subset of the Piccoli-Rossi constraint set, $\Gamma^{\text{EMD}*}_{\leq(\mathcal{E},\tilde{\mathcal{E}})} \subseteq \Gamma_{\leq(\mathcal{E},\tilde{\mathcal{E}})}$. Furthermore, if $\gamma_{ij} \in \Gamma^{\text{EMD}*}_{\leq(\mathcal{E},\tilde{\mathcal{E}})}$, then the values of the objective function in each minimization problem coincide (as $p = 1$ now), up to a factor of $\kappa = 2R$. Thus, if we can show that $\kappa = 2R \geq \max_{ij} d_{ij}$ ensures that the optimizer γ^*_{ij} of T_1^κ belongs to the stricter constraint set $\Gamma^{\text{EMD}*}_{\leq(\mathcal{E},\tilde{\mathcal{E}})}$, we can conclude that

$$T_1^\kappa(\mathcal{E},\tilde{\mathcal{E}}) = T_1^{2R}(\mathcal{E},\tilde{\mathcal{E}}) = R\ \text{EMD}^{*,R}(\mathcal{E},\tilde{\mathcal{E}}). \tag{2.15}$$

Observe that, using properties (2, 3) of the constraint set $\Gamma_{\leq(\mathcal{E},\tilde{\mathcal{E}})}$, we may remove the absolute value signs in the definition of T_1^κ and express it equivalently as

$$\begin{aligned} &T_1^\kappa(\mathcal{E},\tilde{\mathcal{E}}) = \\ &\qquad \min_{\gamma_{ij}\in\Gamma_{\leq(\mathcal{E},\tilde{\mathcal{E}})}} \sum_{ij} (d_{ij} - \kappa)\, \gamma_{ij} + \frac{\kappa}{2}\left(\Sigma_i E_i + \Sigma_j \tilde{E}_j\right) \end{aligned} \tag{2.16}$$

Thus, if $\kappa \geq \max_{ij} d_{ij}$, the coefficient on γ_{ij} is always negative, so the optimal γ^*_{ij} for the T_1^κ distance will be as large as possible, subject to the constraints (2, 3). In particular, the optimal γ^*_{ij} will satisfy constraint (4) and belong to $\Gamma^{\text{EMD}*}_{\leq(\mathcal{E},\tilde{\mathcal{E}})}$.

The above argument not only establishes the equivalence between T_1^κ and $\text{EMD}^{*,R}$ for $\kappa = 2R \geq \max_{ij} d_{ij}$, but also sheds light on the role of the parameter $\kappa > 0$. From Equation (2.14), we observe that smaller κ makes creation and destruction cheaper and transport comparatively more expensive. In fact, using Equation (2.16), we can make this quantitative: if γ^*_{ij} is the optimizer, then for any i, j such that $d_{ij} > \kappa$, we must have $\gamma^*_{ij} = 0$. If not, we could find a strictly better choice of γ in $\Gamma_{\leq(\mathcal{E},\tilde{\mathcal{E}})}$ by setting $\gamma_{ij} = 0$, contradicting that γ^*_{ij} was the optimizer.

In other words, energy will never be transported over a distance greater than κ. This κ parameter sets an intrinsic length scale to any unbalanced optimal transport distance formulated in the above way.

2.2.2 From Partial Optimal Transport to Unbalanced Optimal Transport

One of the key contributions of Piccoli and Rossi's work on the partial optimal transport distance T^κ_p is a *dynamic*, Benamou–Brenier-type formulation of the distance [50]. This dynamic perspective is most clear when T^κ_p is stated in full generality, as a distance on the space of general measures $\mathcal{M}(\Omega)$.

For $\mu_0, \mu_1 \in \mathcal{M}(\Omega)$, $\kappa > 0$, and $p \geq 1$, we have

$$\begin{aligned} &T^\kappa_p(\mu_0, \mu_1) \\ &= \inf_{\gamma \in \Gamma_{\leq(\mu_0,\mu_1)}} \left(\iint |x_0 - x_1|^p d\gamma(x_0, x_1) \right)^{1/p} + \frac{\kappa}{2} \left(\left| \int \mu_0 - \iint \gamma \right| + \left| \int \mu_1 - \iint \gamma \right| \right), \end{aligned} \tag{2.17}$$

where we say $\gamma \in \Gamma_{\leq(\mu_0,\mu_1)}$ in case $\gamma \in \mathcal{M}(\Omega \times \Omega)$ satisfies $\gamma(B \times \Omega) \leq \mu_0(B)$ and $\gamma(\Omega \times B) \leq \mu_1(B)$ for any Borel set B. Note that Equation (2.17) reduces to Equation (2.14) in the discrete case when $\mu_0 = \mathcal{E} = \sum_{i \in I} \delta_{x_i} E_i$ and $\mu_1 = \tilde{\mathcal{E}} = \sum_{j \in J} \delta_{\tilde{x}_j} \tilde{E}_j$.

Piccoli and Rossi [53, 50] showed that T^κ has the following equivalent dynamic formulation,

$$\begin{aligned} T_p^\kappa(\mu_0,\mu_1) &= \inf_{\rho,v,\psi\in\mathcal{CES}(\mu_0,\mu_1)} (A_p^\kappa[\rho,v,\psi])^{1/p}, \qquad (2.18)\\ A_p^\kappa[\rho,v,\psi] &= \int_0^1\int_\Omega (|v(x,t)|^p + (\kappa/2)|\psi(x,t)|)\rho(x,t)\mathrm{d}x\mathrm{d}t,\\ \mathcal{CES}(\mu_0,\mu_1) &= \\ &\{\rho\in C([0,1],\mathcal{M}(\Omega)), v\in L^2(\mathrm{d}\rho_t\mathrm{d}t), \psi\in L^1(\mathrm{d}\rho_t\mathrm{d}t) :\\ &\quad \partial_t\rho + \nabla\cdot(\rho v) = \psi\rho, \rho(\cdot,0)=\mu_0,\ \rho(\cdot,1)=\mu_1\}. \end{aligned}$$

In other words, one can find the T_p^κ distance from μ_0 to μ_1 by considering all curves ρ connecting μ_0 to μ_1 with velocity v and reaction rate ψ and finding the curve with least action $A_p^\kappa[\rho,v,\psi]$.

This dynamic perspective reveals a general framework for unbalanced optimal transport problems, in terms of minimizing different notions of action. In particular, as observed in [53], for any $\kappa>0$, $p\geq 1$, and $q\geq 1$, one may consider

$$A_{p,q}^\kappa[\rho,v,\psi] = \int_0^1\int_\Omega (|v(x,t)|^p + (\kappa/2)^q|\psi(x,t)|^q)\rho(x,t)\mathrm{d}x\mathrm{d}t. \qquad (2.19)$$

As before, large values of $\kappa>0$ penalize creation and destruction. In particular, sending $\kappa\to+\infty$ [52, Theorem 7.24],

$$\begin{aligned} &\lim_{\kappa\to+\infty}\inf_{\rho,v,\psi\in\mathcal{CES}(\mu_0,\mu_1)} \left(A_{p,q}^\kappa[\rho,v,\psi]\right)^{1/p}\\ &\qquad = \begin{cases} \mathrm{W}_p(\mu_0,\mu_1) & \text{if } \int\mu_0 = \int\mu_1\\ +\infty & \text{otherwise.}\end{cases} \qquad (2.20)\end{aligned}$$

While minimizing the action $A_{p,q}^\kappa[\rho,v,\psi]$ with $q=1$ yields the partial transport distance T_p^κ described in the previous section, minimizing it for $p = q = 2$ yields the Hellinger–Kantorovich distance,

$$\mathrm{HK}_\kappa(\mu_0, \mu_1) = \inf_{\rho,v,\psi \in \mathcal{CES}(\mu_0,\mu_1)} (A_{2,2}^\kappa[\rho, v, \psi])^{1/2}. \tag{2.21}$$

This case is distinguished among all $p, q \geq 1$, since it is the only choice that directly gives rise to an infinite dimensional Riemannian manifold [53, 55], although as of now this structure is still much less understood than the W_2 case. We now describe this Hellinger-Kantorovich distance in more details.

2.2.3 HK Distance: Benamou–Brenier-type Formulation

To make the parallelism between HK and W_2 more manifest, we now rewrite Equation (2.18) in the specific case of the HK distance using the same terminology as the corresponding W_2 descriptions.

Essentially, one adds an additional source term on $[0,1]\times\Omega$ to the continuity equation, whose solutions are now denoted by the set $\mathcal{CES}(\mu_0, \mu_1)$ which composes of triplets of measures $(\rho, \omega, \zeta) \in \mathcal{M}([0,1] \times \Omega)^{1+d+1}$ where ρ interpolates between μ_0 and μ_1 and that solve

$$\partial_t \rho + \nabla \omega = \zeta \tag{2.22}$$

in a distributional sense. Here we redefine $\omega := \rho v$ and $\zeta := \rho\psi$ for brevity and consistency of notation. More precisely, we require for all $\phi \in \mathrm{C}^1([0,1] \times \Omega)$ that

$$\int_{[0,1]\times\Omega} \partial_t \phi \,\mathrm{d}\rho + \int_{[0,1]\times\Omega} \nabla\phi \cdot \mathrm{d}\omega + \int_{[0,1]\times\Omega} \phi \,\mathrm{d}\zeta = \int_\Omega \phi(1,\cdot)\,\mathrm{d}\mu_1 - \int_\Omega \phi(0,\cdot)\,\mathrm{d}\mu_0. \tag{2.23}$$

As in the case of W_2, let $J_{\mathrm{HK}} : \mathcal{M}([0,1]\times\Omega)^{1+d+1} \to \mathbb{R}\cup\{\infty\}$ be given by

$$J_{\mathrm{HK},\kappa}(\rho,\omega,\zeta) := \begin{cases} \int_{[0,1]\times\Omega} \left(\left\| \frac{\mathrm{d}\omega}{\mathrm{d}\rho} \right\|^2 + \frac{\kappa^2}{4} \left(\frac{\mathrm{d}\zeta}{\mathrm{d}\rho} \right)^2 \right) \mathrm{d}\rho & \text{if } \rho \geq 0, \omega, \zeta \ll \rho, \\ +\infty & \text{else.} \end{cases} \tag{2.24a}$$

Then for $\mu_0, \mu_1 \in \mathcal{M}_+(\Omega)$ we set

$$\mathrm{HK}_\kappa(\mu_0,\mu_1)^2 := \inf\left\{ J_{\mathrm{HK},\kappa}(\rho,\omega,\zeta) | (\rho,\omega,\zeta) \in \mathcal{CES}(\mu_0,\mu_1) \right\}. \tag{2.24b}$$

The similarity between HK and W_2 is now obvious from this reformulation.

Again, the parameter $\kappa > 0$ controls the relative importance of the transport part of the cost—the first term $\int_{[0,1]\times\Omega} \|\frac{\mathrm{d}\omega}{\mathrm{d}\rho}\|^2 \mathrm{d}\rho$, and the destruction/creation part—the second term $\int_{[0,1]\times\Omega} (\frac{\mathrm{d}\zeta}{\mathrm{d}\rho})^2 \mathrm{d}\rho$. As in the general case of $T^\kappa_{p,q}$, the HK metric admits a well defined limit as $\kappa \to +\infty$ whenever μ_0 and μ_1 have equal mass, i.e., $\lim_{\kappa\to+\infty} \mathrm{HK}_\kappa(\mu_0,\mu_1) = \mathrm{W}_2(\mu_0,\mu_1)$.

On the other hand, the $\kappa \to 0$ limit is also well defined for arbitrary μ_0, μ_1,

$$\lim_{\kappa\to 0} \frac{1}{\kappa} \mathrm{HK}_\kappa(\mu_0,\mu_1) = \left(\int \left| \sqrt{\frac{\mathrm{d}\mu_0}{dx}} - \sqrt{\frac{\mathrm{d}\mu_1}{\mathrm{d}x}} \right|^2 \mathrm{d}x \right)^{1/2}, \tag{2.25}$$

which is known as the *Hellinger distance* [52, Theorems 7.22 and 7.24].

Seen from a different angle, setting the parameter κ is equivalent to re-scaling the set Ω to Ω/κ (and all measures accordingly), then computing HK_1 on Ω/κ and finally multiplying the result by κ again. Transport in HK_1 is bounded by $\frac{\pi}{2}$, in the sense that mass at location x_0 will never be transported outside of the ball $B(x_0, \pi/2)$. Therefore the transport in HK_κ is bounded by $\frac{\kappa\pi}{2}$.

This observation is useful in applications as it allows one to prescribe how far mass should be transported and provides a good intuition for the choice of κ. Usually, as we will later confirm, statistical performance with respect to κ is relatively robust around the optimal value. Therefore, a coarse cross validation search for an acceptable value is oftentimes sufficient.

It can be proven [53, Theorems 2.1 and 2.2] that HK is a well-defined metric on $\mathcal{M}_+(\Omega)$. Minimizers of (2.24b) exist and will again be referred to as *constant speed geodesics* between μ_0 and μ_1 with respect to HK.

2.2.4 HK Distance: Kantorovich-type Formulation

Similar to the classical W_2 distance, there are also (multiple) static, *Kantorovich-type* formulations [52] of HK in terms of measures on the product space $\Omega \times \Omega$, which lends itself to familiar numerical approximations, such as via entropic regularization. Unlike in the classical balanced case, here transport can no longer be described by a coupling $\pi \in \Pi(\mu_0, \mu_1)$, since particles may change their mass during transport and μ_0 and μ_1 may have different total mass.

In the static formulation for the Hellinger–Kantorovich given in [52], the effect of mass changes is captured by choosing a particular cost function and by relaxing the marginal constraints $P_{i\sharp}\pi = \mu_i$ and penalizing the difference with the Kullback–Leibler divergence instead. The Kullback–Leibler divergence of $\mu \in \mathcal{M}(\Omega)$ with respect to $\nu \in \mathcal{M}(\Omega)$ is defined to be

$$\mathrm{KL}(\mu|\nu) = \begin{cases} \int \varphi(\frac{\mathrm{d}\mu}{\mathrm{d}\nu})\,\mathrm{d}\nu & \text{if } \mu, \nu \geq 0, \mu \ll \nu, \\ +\infty & \text{else} \end{cases} \tag{2.26}$$

with $\varphi(s) = s\log(s) - s + 1$ for $s > 0$ and $\varphi(0) = 1$, which is is strictly convex and continuous on $\mathbb{R}_+$. Note that the KL divergence is in general not symmetric with respect to $\mu \leftrightarrow \nu$.

The Kantorovich formulation of HK can now be expressed via a particular "soft-marginal" Kantorovich-type transport problem [52, Theorem 8.18]. Let

$$c(x_0, x_1) := \begin{cases} -2\log(\cos(\|x_0 - x_1\|)) & \text{if } \|x_0 - x_1\| < \frac{\pi}{2} \\ +\infty & \text{else.} \end{cases} \tag{2.27a}$$

$$J_{\mathrm{SM}}(\pi) := \int_{\Omega^2} c \, \mathrm{d}\pi + \sum_{i \in \{0,1\}} \mathrm{KL}(P_{i\sharp}\pi | \mu_i). \tag{2.27b}$$

Then

$$\mathrm{HK}(\mu_0, \mu_1)^2 = \inf \left\{ J_{\mathrm{SM}}(\pi) \,\middle|\, \pi \in \mathcal{M}_+(\Omega^2) \right\} \tag{2.27c}$$

and minimal π in (2.27c) exist. Here we set $\kappa = 1$ for simplicity.

As in the Wasserstein case, there are conditions under which a unique optimal transport map exist for HK. Let $\mu_0 \in \mathcal{M}_{+,\mathcal{L}}(\Omega)$, $\mu_1 \in \mathcal{M}_+(\Omega)$. Then the minimizer π for $\mathrm{HK}(\mu_0, \mu_1)^2$ in (2.27) is unique and induced by a Monge map. This is the same as in the W_2 case.

The Special Case of Discrete Measures

For ease of use in latter physics applications, we now rephrase the above static minimization problem of the HK distance in the special case of fully discrete measures and highlight the intuition behind the mathematical formulation.

For two discrete measures, $\mathcal{E} = \sum_{i \in I} \delta_{x_i} E_i$ and $\tilde{\mathcal{E}} = \sum_{j \in J} \delta_{\tilde{x}_j} \tilde{E}_j$, the HK distance is given by

$$\begin{aligned}&\mathrm{HK}_\kappa(\mathcal{E},\tilde{\mathcal{E}}) \qquad (2.28)\\ &= \min_{\gamma_{ij}\geq 0}\sum_{ij}\Big(\ell^\kappa(d_{ij})\gamma_{ij}+\kappa^2\mathrm{KL}(\mathcal{G},\mathcal{E})+\kappa^2\mathrm{KL}(\tilde{\mathcal{G}},\tilde{\mathcal{E}})\Big)^{1/2},\end{aligned}$$

where $\mathcal{G}$ and $\tilde{\mathcal{G}}$ are auxiliary discrete measures, with $\mathcal{G}$ assigning mass $G_i = \sum_j \gamma_{ij}$ to location x_i and $\tilde{\mathcal{G}}$ assigning mass $\tilde{G}_j = \sum_i \gamma_{ij}$ to $\tilde{x}_j$. In addition, we define

$$\begin{aligned}\ell^\kappa(s) &= \begin{cases}-2\kappa^2\log(\cos^2(s/\kappa)) & \text{if } s<\frac{\pi}{2}\kappa,\\ +\infty, & \text{otherwise,}\end{cases} \qquad (2.29)\\ \mathrm{KL}(\mathcal{G},\mathcal{E}) &= \sum_i E_i\varphi\left(\frac{G_i}{E_i}\right), \quad \varphi(s)=s\log(s)-s+1.\end{aligned}$$

The equivalence between Equation (2.21) and Equation (2.28), i.e., the dynamic and static formulations, is a significant mathematical result, due to Liero, Mielke, and Savaré, based on a surprising connection with cone geometry [51, 52].

The optimizer γ_{ij} of Equation (2.28) represents how much mass is transported from x_i in $\mathcal{G}$ to $\tilde{x}_j$ in $\tilde{\mathcal{G}}$. That is, γ_{ij} is the optimal transport plan from $\mathcal{G}$ to $\tilde{\mathcal{G}}$. In general, $G_i \neq E_i$ and $\tilde{G}_j \neq \tilde{E}_j$, and the mass that is not transported can be thought of as having been created or destroyed. In particular,

1. if $G_i > E_i$, we say energy was *created* at x_i;
2. if $G_i < E_i$, we say energy was *destroyed* at x_i;
3. if $G_j > \tilde{E}_j$, we say energy was *destroyed* at $\tilde{x}_j$;
4. if $G_j < \tilde{E}_j$, we say energy was *created* at $\tilde{x}_j$.

Note that the first and third options did not arise for the T_p^κ distance, due to requirements

(2, 3) for the set of transport plans $\Gamma_{\leq(\mathcal{E},\tilde{\mathcal{E}})}$. While until now we have always assumed that our discrete measures have strictly positive energy at every location, $E_i, \tilde{E}_j > 0$, observe that now it is possible for G_i or $\tilde{G}_j$ to be zero.

Again, the first term in the minimization problem in Equation (2.28) penalizes transporting energy over long distances. As with T_p^κ, small values of κ penalize transport. The second two terms penalize the difference between $\mathcal{G}$ and $\mathcal{E}$ and between $\tilde{\mathcal{G}}$ and $\tilde{\mathcal{E}}$, in terms of the Kullback-Liebler divergence.

As we all know, the major difference between the Hellinger–Kantorovich metric and the 2-Wasserstein metric is that HK allows for the comparison of events with unequal total energy. However, even when the total energy of events $\mathcal{E}$ and $\tilde{\mathcal{E}}$ coincide, $\mathrm{HK}_\kappa(\mathcal{E},\tilde{\mathcal{E}})$ is in general not equal to $\mathrm{W}_2(\mathcal{E},\tilde{\mathcal{E}})$. This can be seen, for example, from Equation (2.28) and Equation (2.29): mass will never be transported more than distance $\frac{\kappa\pi}{2}$. Interestingly, the converse is also true. If mass is not transported from x_i to $\tilde{x}_j$, i.e., if $\gamma_{ij} = 0$, then we must have $d_{ij} = \|x_i - \tilde{x}_j\| \geq \frac{\kappa\pi}{2}$ [34, Lemma 3.13].

2.2.5 Global Mass Rescaling Behavior of HK

The introduction of unbalanced optimal transport, specifically the HK distance, has originally been to accommodate cases where the two measures being compared have unequal total masses. From the formulation of HK above, we see that HK accomplishes this extension by allowing mass to be created and destroyed locally at each location. The question now arises: is the global mass difference between two measures more important, or is it the local mass discrepancy that distinguishes HK from W_2?

The answer favors the latter. Specifically, let $\mu_0, \mu_1 \in \mathcal{M}_1(\Omega)$, i.e., both with total mass equal to unity, and let $m_0, m_1 \in \mathbb{R}_+$ represent the respective scaling factor for the total mass of each measure. It was shown in [63, Theorem 3.3] that

$$\mathrm{HK}(m_0 \cdot \mu_0, m_1 \cdot \mu_1)^2 = \sqrt{m_0 \cdot m_1} \cdot \mathrm{HK}(\mu_0, \mu_1)^2 + (\sqrt{m_0} - \sqrt{m_1})^2 \tag{2.30}$$

and if π is optimal in Equation (2.27) for $\mathrm{HK}(\mu_0, \mu_1)^2$, then $\sqrt{m_0 \cdot m_1} \cdot \pi$ is optimal for $\mathrm{HK}(m_0 \cdot \mu_0, m_1 \cdot \mu_1)^2$.

This is a significant result. What it implies is that the "unbalanced" effects of the HK distance are already fully encoded in its behavior on probability measures, i.e., balanced measures with unit total mass. Extension to measures of arbitrary mass can then be done via the simple formula above. Consequently, the benefit for data analysis applications that we expect from using HK instead of W_2 is not so much the ability to deal with differences in the *total* mass of measures, but its ability to deal with *local* mass discrepancies, i.e., creating mass in one part of the distribution while reducing it in another part, if this seems more likely than a long range transport.

Therefore, for numerical purposes we always normalize our samples before comparison, as will be done for all our later applications. In the case where the total mass of samples is deemed relevant for the subsequent analysis, its effect can be easily recovered via Equation (2.30). Or the total masses can also be kept as separate features, though this later option has not been pursued in our current study.

2.2.6 Geodesics for HK

Again, we consider the HK distance between two Dirac measures, where the general idea is outlined without proofs. For more details, please refer to Section 3 in [34].

Let $x_0, x_1 \in \Omega$, $m_0, m_1 \in \mathbb{R}_+$, producing the two Dirac measures with different total masses, i.e., $\mathcal{E}_0 = \delta_{x_0} \cdot m_0$ and $\mathcal{E}_1 = \delta_{x_1} \cdot m_1$. The HK distance between them are given

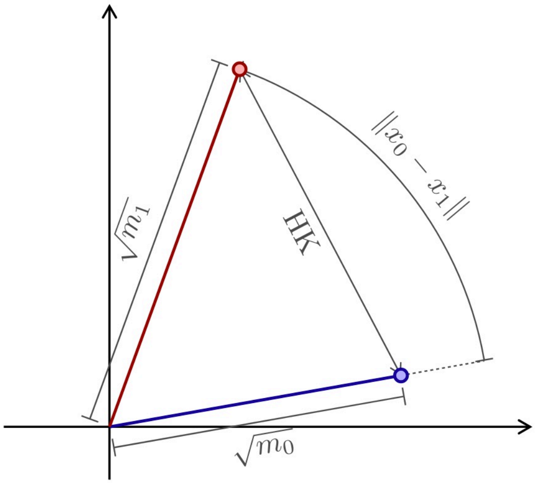


Figure 2.2: Local isometry between Dirac measures $\delta_{x_i} \cdot m_i$, $i = 0, 1$, with respect to the HK metric and points in $\mathbb{C}$ with respect to the Euclidean distance. When $\|x_0 - x_1\| \leq \frac{\pi}{2}$, the geodesic between the two measures is described by the corresponding straight line in $\mathbb{C}$. Figure copied from [34].

by [52, 53]

$$\mathrm{HK}(\delta_{x_0} \cdot m_0, \delta_{x_1} \cdot m_1)^2 = m_0 + m_1 - 2\sqrt{m_0\, m_1}\overline{\cos}(\|x_0 - x_1\|) \tag{2.31}$$

where $\overline{\cos}(s) = \cos\left(\min\{s, \frac{\pi}{2}\}\right)$.

An intuitive visualization for Equation (2.31) is given in Figure 2.2. For $\|x_0 - x_1\| \leq \frac{\pi}{2}$, the HK distance equals the distance between two points in $\mathbb{C}$ in polar coordinates, where the radii of the two vectors are given by $\sqrt{m_i}$ respectively and the angle between them is $\|x_0 - x_1\|$. That is, one can write

$$\mathrm{HK}(\delta_{x_0} \cdot m_0, \delta_{x_1} \cdot m_1) = \|\sqrt{m_0} - \sqrt{m_1}\exp(i\|x_0 - x_1\|)\| \tag{2.32}$$

From this local isometry, one can deduce the structure of geodesics between Dirac measures. If $\|x_0 - x_1\| < \frac{\pi}{2}$, the geodesic is given by a single “travelling” Dirac of the form

$$\rho_t = \delta_{X(t)} \cdot M(t) \tag{2.33}$$

where $X : [0,1] \to \Omega$ describes the movement from x_0 to x_1 and $M : [0,1] \to \mathbb{R}_+$ the evolution of the mass. For W_2, one would of course have that $X(t)$ parametrizes the constant speed straight line from x_0 to x_1 and $M(t)$ would be fixed to 1. On the other hand, for the HK metric, X and M essentially describe the straight line between the two embedded points $\sqrt{m_0}$ and $\sqrt{m_1}\exp(i\|x_0 - x_1\|)$ in $\mathbb{C}$ in polar coordinates, with the angle given by X and the squared radius given by M.

When $\|x_0 - x_1\| > \frac{\pi}{2}$, the geodesic between the two measures is equal to the geodesic in the Hellinger distance: the mass at x_0 is decreased from m_0 to 0, the mass at x_1 is increased from 0 to m_1 and no transport occurs. The geodesic will be of the form

$$\rho_t = \delta_{x_0} \cdot M_0(t) + \delta_{x_1} \cdot M_1(t)\,. \tag{2.34}$$

Explicit formulas for X and M, for M_0 and M_1, and for $(\rho,\, \omega,\, \zeta)$, are given in [34, Proposition 3.7]. A similar illustration of the Dirac-to-Dirac geodesics as in the W_2 distance is presented in Figure 2.3.

As in the 2-Wasserstein distance, constant speed geodesics of $\mathrm{HK}(\mu_0, \mu_1)$ for general measures can also be constructed via the superposition of Dirac-to-Dirac geodesics. The formula is more complicated here; see Section 3.3 in [34] for more details.

2.3 Computational Hurdle of Optimal Transport

In the previous two sections, we have seen the theoretical benefits of optimal transport metrics, in particular, the balanced W_2 distance and the unbalanced HK distance. Both

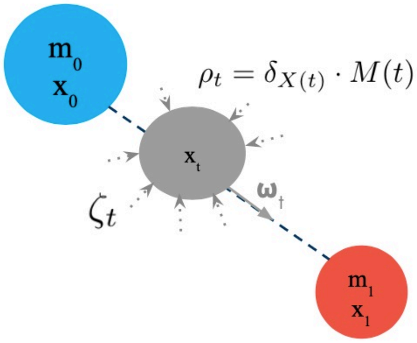


Figure 2.3: An illustration for the Dirac-to-Dirac geodesic in HK.

come with a rich geometric structure, such as geodesics, barycenters [64], and a weak Riemannian structure to be explained later. On the practical side, as discussed in Chapter 1, the OT distances often better capture variations in signals and images comparing to "pointwise" similarity measures such as Euclidean ℓ^p-norms or the Kullback–Leibler divergence.

However, the wider adoption of OT in image analysis, and more broadly in any applied science, has been slowed by two main obstacles: high computational cost and limited choice of downstream statistical analysis models. In terms of computational efficiency, despite recent numerical advances (see [65] for an overview), it is still relatively expensive to compute OT distances, particularly for large and high dimensional data sets. For example, computing the balanced p-Wasserstein distance between two discrete measures, each with n Dirac masses, requires $O(n^3)$ operations via Bertsekas' auction algorithm and $O(n^2 \log(n))$ operations via entropic regularization and the Sinkhorn algorithm [66, 67, 68, 69, 37]. This is in stark contrast to the classical ℓ^2 norm, which is naively $\mathcal{O}(n)$, when the number of bins is chosen proportional to the number of Diracs.

The high cost of the OT metrics is compounded by the fact that one needs to compute the pairwise distances between the entire collection of N data points, requiring $O(N^2)$ computations of the OT distance itself. In the particular case of classifying collider jets considered in Chapter 3, the number of particles per jet is relatively small, i.e., $n \approx 10^2$. It is therefore the latter need to compute pairwise distances between a large number of jets, $N \approx 10^5$, which is the main computational expense.

To give a sense of time scale we are talking about, computing the W_2 distance between two jets, each with $\mathcal{O}(100)$ particles, takes fractions of a second, whereas the normal Euclidean calculation only takes milliseconds. This quickly puts the calculation of W_2 distances for a typical dataset with $\mathcal{O}(10^5)$ jets beyond the reach of desktop computers. The generalization to the unbalanced case only makes things worse. This computational hurdle poses a serious challenge to the usability of optimal transport distances for our physics analyses.

Furthermore, existing work using classical optimal transport metrics must also cope with the significant computational demands of storing the matrix of pairwise distances, which is itself oftentimes unsuitable for use with downstream machine learning methods that require more structure than just the pairwise distances. Therefore, when attempting to perform a given statistical task on an input dataset, we encounter the second issue of having fewer models available for analyzing the OT distances themselves.

Fortunately, we can kill two birds with one stone. Below, we introduce the Linearized Optimal Transport (LOT) approximation to the exact OT distances. We first give the general principle for the LOT framework and highlight some high-level properties. The following two sections then expound the idea specifically for the 2-Wasserstein distance and the Hellinger-Kantorovich distance and work out the necessary mathematical details. Note that the development of the LOT framework is one of the main contributions of the three publications by the author and the collaborators [35, 34, 36].

2.3.1 Linearized Optimal Transport

Thanks to their weak Riemannian structure, it is possible to linearize the two special OT distances, W_2 and HK, which is why we have been focusing on them in the first place. Formally, this means that the W_2/HK manifold can be approximated locally by a tangent space at a reference point. One then apply a logarithmic map, which brings the data samples on the original OT manifold down to the tangent space. Later we will see that the tangent space is a Hilbert space equipped with an inner product, where the norm of each embedded sample equals its W_2/HK distance to the reference measure. The good news is that the Hilbertian distance between two embedded samples is also approximately equal to the W_2/HK distance between the two original samples.

In other words, the LOT approximation to the exact OT distance amounts to projecting everything onto the tangent plane at a chosen reference event and computing simpler ℓ^2 distances on that plane. This makes the computational advantage of the LOT approach very obvious. Now instead of calculating $\mathcal{O}(N^2)$ computationally intensive OT distances between each pair of data samples (with a total of N samples), one only needs to solve N optimal transport problems in order to embed N samples into the tangent space at the reference point. The pairwise LOT distances are then obtained by $\mathcal{O}(N^2)$ computationally efficient ℓ^2 distances. In practice, this linear version reduces the computational effort from a computer cluster to a single PC.

Another advantage of LOT lies in the linear structure itself on the space of embeddings, in comparison to the non-linear metric structure of the original W_2/HK space. Such a Euclidean embedding allows for the application of methods from data analysis that have been primarily developed for linear settings such as principal component analysis (PCA), opening the door for a wide range of machine learning algorithms. On the other hand, since the obtained embedding is (at least locally) approximating the original W_2/HK distance, we expect the tangent linear structure to still enjoy many of the properties of the W_2/HK distance such as the cost of translations, as opposed to other naive linear structures on the space of measures.

The linearization of the 2-Wasserstein distance was proposed in [70] and recent work by Delalande and Mérigot [71] has quantified the relationship between the original 2-Wasserstein distance and its linearization. Applications in relatively simple settings have emerged over the past decade. In [70], linearized W_2 was employed as a method for visualizing variation in sets of images. [72] used the linear setting to generate new images by first clustering in the linear W_2 space, then learning the principal directions in the tangent planes for each cluster. Hence, new images were generated using Euclidean data analysis techniques, such as k-means and PCA, whilst keeping the Wasserstein flavor. Other applications of the linear 2-Wasserstein space have included a PCA based approach for super resolution on faces [73], and classification, using a Fisher linear discriminant analysis technique, on images of nuclei [74].

Comparing to the case of the linearized W_2 where a tangent vector is represented by a velocity field (see Section 2.4 for more details), when linearizing the HK distance, one obtains an additional scalar mass creation/destruction field (see Section 2.5). This field can become singular in the case where mass is created from nothing, leading to a third, measure-valued, tangent component. This third component may be considered undesirable in some applications and sufficient conditions can be imposed to ensure that it remains zero. This way, one obtains the desired embedding into a Hilbert space.

Although the formalism of the linearized HK distance may seem considerably more complex compared to the W_2 case, from a numerical perspective the local linearization via HK is not much harder to perform than for W_2. The involved transport problems can be solved in a Kantorovich-type formulation, for instance with an adapted Sinkhorn algorithm, just like for W_2. Then an approximate logarithmic map can be extracted from the optimal coupling with explicit formulas, leading to an embedding into the Hilbert space, which becomes finite-dimensional after discretization. Similar numerical approximations as in the case of W_2 apply. The only new challenge is to fix the intrinsic length-scale κ of the HK metric appropriately, which can be determined by standard validation procedures.

We now describe in detail the Riemannian structure of the two special optimal transport distances, leading to their linearization on the space of general measures. A proper choice of a reference measure is critical and ensures the resulting LOT distances are genuine metrics themselves. In practice, however, one usually has only discrete measures, both for the input data and for the reference. Therefore, this discrete case is treated separately, where the linearized W_2 and HK are sometimes referred to as pseudo-distances to highlight the fact that they are themselves approximation to the LOT distances in the continuous setting.

2.4 Linearized 2-Wasserstein Metric

In this section, we describe the procedure to linearize the 2-Wasserstein distance, with the resulting distance shorthanded as "LinW_2". We first briefly explain the Riemannian structure of the W_2 distance and work out the corresponding logarithmic and exponential maps. The logarithmic map in specific gives us the linearization scheme, which under certain condition of the reference measure outputs a true metric on the space of general measures.

We then zoom in to the setting of discrete measures more relevant for practical use. It is this formalism that will be adopted in our physics analysis. As now the reference measure no longer satisfies the required condition, the resulting linearization in general does not give a metric, but an approximation to the true LinW_2 metric. We denote this *pseudo-distance* as "$\mathrm{LinW}_{2,\mathcal{R}}$", with the added subscript $\mathcal{R}$ emphasizing the discrete nature of the reference measure. Finally, we include a simple example of two Dirac distributions (also called "artificial jets" alluding to the latter collider application) to illustrate the actual implementation of the $\mathrm{LinW}_{2,\mathcal{R}}$ framework.

2.4.1 Riemannian structure of W$_2$

As pointed out earlier, the 2-Wasserstein metric is special among all balanced p-Wasserstein metrics because only it enjoys a Riemannian structure. Here we give an intuitive explanation that will facilitate the latter linearization of the distance. A more complete picture with rigorous proofs can be found, for instance, in [75, Sections 2.3.2 and 7.2].

Equations (2.5) (2.6), at the formal level, look like a functional to find constant speed geodesics on a Riemannian manifold. Here the manifold is $\mathcal{M}_1(\Omega)$, the curve is given by $t \mapsto \rho_t$, and the tangent vectors are encoded by the velocity field $v_t := \frac{\mathrm{d}\omega_t}{\mathrm{d}\rho_t}$. The Riemannian inner product between tangent vectors v and w at ρ_t is then given by

$$g_{\mathrm{W}_2}(\rho_t; v, w) := \int_\Omega \langle v, w \rangle \, \mathrm{d}\rho_t. \tag{2.35}$$

The relation between tangent vectors v_t and the curve ρ_t is encoded in the continuity equation.

Now assume additionally that $\mu_0 \in \mathcal{M}_{1,\mathcal{L}}(\Omega)$, i.e. $\mu_0 \ll \mathcal{L}$. In this case, we know there exists a unique optimal coupling $\pi \in \Pi(\mu_0, \mu_1)$ for $\mathrm{W}_2(\mu_0, \mu_1)$. Let (ρ, ω) be the corresponding geodesics constructed via Equations (2.11) (2.12). One then finds that

$$\mathrm{W}_2(\mu_0, \mu_1)^2 = \int_0^1 \int_\Omega \|v_t\|^2 \, \mathrm{d}\rho_t \, \mathrm{d}t \qquad \text{with} \qquad v_t := \frac{\mathrm{d}\omega_t}{\mathrm{d}\rho_t}. \tag{2.36}$$

Thanks to the assumption on μ_0, the unique optimal plan π is induced by a Monge map $\mathbf{t} : \Omega \to \Omega$, i.e., $\pi = (\mathbf{id}, \mathbf{t})_\sharp \mu_0$. In this particular case, we have

$$\rho_t = X(\cdot, \cdot; t)_\sharp (\mathbf{id}, \mathbf{t})_\sharp \mu_0 = \big((1-t) \cdot \mathbf{id} + t \cdot \mathbf{t}\big)_\sharp \mu_0, \tag{2.37}$$

$$\omega_t = X(\cdot, \cdot; t)_\sharp \left(\partial_t X(\cdot, \cdot; t) \cdot (\mathbf{id}, \mathbf{t})_\sharp \mu_0\right) = \big((1-t) \cdot \mathbf{id} + t \cdot \mathbf{t}\big)_\sharp \big((\mathbf{t} - \mathbf{id}) \cdot \mu_0\big) \tag{2.38}$$

and thus $v_t\big((1-t)\, x_0 + t\, \mathbf{t}(x_0)\big) = \mathbf{t}(x_0) - x_0$. In particular,

$$v_0(x_0) = \mathbf{t}(x_0) - x_0. \tag{2.39}$$

Consequently, for all $t \in [0, 1]$

$$\begin{aligned} \int_\Omega \left\| \frac{\mathrm{d}\omega_t}{\mathrm{d}\rho_t} \right\|^2 \mathrm{d}\rho_t &= \int_\Omega \|v_t\|^2 \, \mathrm{d}\big((1-t) \cdot \mathbf{id} + t \cdot \mathbf{t}\big)_\sharp \mu_0 \\ &= \int_\Omega \big\|v_t \circ \big((1-t) \cdot \mathbf{id} + t \cdot \mathbf{t}\big)\big\|^2 \, \mathrm{d}\mu_0 \\ &= \int_\Omega \|v_0\|^2 \, \mathrm{d}\mu_0. \end{aligned} \tag{2.40}$$

Finally,

$$\mathrm{W}_2(\mu_0, \mu_1)^2 = \int_\Omega \|v_0\|^2 \, \mathrm{d}\mu_0 = g_{\mathrm{W}_2}(\mu_0; v_0, v_0), \tag{2.41}$$

where we use Equation (2.35) in the last step.

We interpret the map $\mathbf{t} \mapsto v_0$ implied by Equation (2.39) such that it takes μ_1 to the tangent vector at $t = 0$ of the constant-speed geodesic from μ_0 to μ_1. Thus, it is formally the logarithmic map at μ_0 and we will denote it in the following as $\mathrm{Log}_{\mathrm{W}_2}(\mu_0; \cdot)$.

With this, Equation (2.41) becomes

$$\mathrm{W}_2(\mu_0,\mu_1)^2 = g_{\mathrm{W}_2}\big(\mu_0; \mathrm{Log}_{\mathrm{W}_2}(\mu_0;\mu_1), \mathrm{Log}_{\mathrm{W}_2}(\mu_0;\mu_1)\big). \tag{2.42}$$

This is analogous to the classical result in (finite-dimensional) Riemannian geometry. As is well known, the corresponding exponential map is given by $\mathrm{Exp}_{\mathrm{W}_2}(\mu_0, v_0) := (\mathbf{id} + v_0)_\sharp\mu_0$. One finds that $\mathrm{Exp}_{\mathrm{W}_2}(\mu_0, v_0) = \mathbf{t}_\sharp\mu_0 = \mu_1$, as expected.

2.4.2 LinW$_2$ Metric on the Space of Probability Measures

Let's work out the LinW$_2$ metric on the space of general probability measures. Let $\mu_1, \mu_2 \in \mathcal{M}_1(\Omega)$, which can in principle take on a discrete form. We specifically pick the reference measure $\mu_0 \in \mathcal{M}_{1,\mathcal{L}}(\Omega)$, i.e., μ_0 is a Lebesgue-absolutely continuous probability measure. As proposed in [70] for applications in the geometric analysis of ensembles of images, we can use Equation (2.42) to linearize W_2 around the support point μ_0. That is, we define the LinW$_2$ distance as

$$\begin{aligned}\mathrm{LinW}_2(\mu_0;\mu_1,\mu_2)^2 &:= g_{\mathrm{W}_2}\big(\mu_0; \mathrm{Log}_{\mathrm{W}_2}(\mu_0;\mu_1) - \mathrm{Log}_{\mathrm{W}_2}(\mu_0;\mu_2), \\ &\qquad\qquad \mathrm{Log}_{\mathrm{W}_2}(\mu_0;\mu_1) - \mathrm{Log}_{\mathrm{W}_2}(\mu_0;\mu_2)\big) \\ &= \int_\Omega \big\|\mathrm{Log}_{\mathrm{W}_2}(\mu_0;\mu_1) - \mathrm{Log}_{\mathrm{W}_2}(\mu_0;\mu_2)\big\|^2 \,\mathrm{d}\mu_0,\end{aligned} \tag{2.43}$$

where in the last step we again use Equation (2.35). We have now written $\mathrm{LinW}_2(\mu_0;\mu_1,\mu_2)$ as the $\ell^2(\mu_0)$ distance on Ω between $\mathrm{Log}_{\mathrm{W}_2}(\mu_0;\mu_1)$ and $\mathrm{Log}_{\mathrm{W}_2}(\mu_0;\mu_2)$ and can prove that $\mathrm{LinW}_2(\mu_0;\mu_1,\mu_2)$ is indeed a metric.

When the Monge map $\mathbf{t}$, used in the definition of the logarithmic map, does not exist, one instead uses the shortest generalized geodesic (see also [76, Definition 9.2.2]). That is, we let

$$\gamma \in \mathcal{M}_1(\Omega^3) \quad \text{with} \quad P_{i\sharp}\gamma = \mu_i, P_{01\sharp}\gamma \in \Pi_{\mathrm{opt}}(\mu_0, \mu_1) \text{ and } P_{02\sharp}\gamma \in \Pi_{\mathrm{opt}}(\mu_0, \mu_2) \tag{2.44}$$

(such a γ exists by [76, Lemma 5.3.2]) and define the linearized W_2 distance by

$$\mathrm{LinW}_2(\mu_0; \mu_1, \mu_2)^2 := \min_{\gamma \text{ satisfying } (2.44)} \int_{\Omega^3} \|x_1 - x_2\|^2 \, \mathrm{d}\gamma(x_0, x_1, x_2). \tag{2.45}$$

In the case when the Monge map exists, $\Pi_{\mathrm{opt}}(\mu_0, \mu_1)$ and $\Pi_{\mathrm{opt}}(\mu_0, \mu_2)$ then contain a single transport plan, each of which can be written as $\pi_{01} = (\mathbf{id}, \mathbf{t}_0^1)_\sharp \mu_0 \in \Pi_{\mathrm{opt}}(\mu_0, \mu_1)$, $\pi_{02} = (\mathbf{id}, \mathbf{t}_0^2)_\sharp \mu_0 \in \Pi_{\mathrm{opt}}(\mu_0, \mu_2)$. Furthermore, the γ that satisfies (2.44) is unique and given by $\gamma = (\mathbf{id}, \mathbf{t}_0^1, \mathbf{t}_0^2)_\sharp \mu_0$. In this case, Equation (2.43) and Equation (2.45) coincide, as wished.

If the Monge map does not exist, then the minimization in Equation (2.45) is no longer necessarily over a singleton, even though optimal plans π_{01} and π_{02} might still be unique. To remove the minimization in Equation (2.45) in the general case (without Monge maps), following [70], one approximates optimal plans $\pi_{01} \in \Pi_{\mathrm{opt}}(\mu_0, \mu_1)$, $\pi_{02} \in \Pi_{\mathrm{opt}}(\mu_0, \mu_2)$ by a plan induced by a map through barycentric projection, namely

$$\pi_{01} \approx (\mathbf{id}, \mathbf{t}_0^1)_\sharp \mu_0, \qquad \mathbf{t}_0^1(x_0) := \int_\Omega x_1 \, \mathrm{d}\pi_{01,x_0}(x_1), \tag{2.46}$$

$$\pi_{02} \approx (\mathbf{id}, \mathbf{t}_0^2)_\sharp \mu_0, \qquad \mathbf{t}_0^2(x_0) := \int_\Omega x_2 \, \mathrm{d}\pi_{02,x_0}(x_2) \tag{2.47}$$

where $\{\pi_{01,x_0}\}_{x_0 \in \Omega} \subset \mathcal{M}_1(\Omega)$, $\{\pi_{02,x_0}\}_{x_0 \in \Omega} \subset \mathcal{M}_1(\Omega)$ are the disintegrations of π_{01}, π_{02}

with respect to μ_0, i.e.

$$\int_{\Omega^2} \phi(x_0, x_1)\, \mathrm{d}\pi_{01}(x_0, x_1) = \int_{\Omega} \left(\int_{\Omega} \phi(x_0, x_1)\, \mathrm{d}\pi_{01,x_0}(x_1) \right) \mathrm{d}\mu_0(x_0), \tag{2.48}$$

$$\int_{\Omega^2} \phi(x_0, x_2)\, \mathrm{d}\pi_{02}(x_0, x_2) = \int_{\Omega} \left(\int_{\Omega} \phi(x_0, x_2)\, \mathrm{d}\pi_{02,x_0}(x_2) \right) \mathrm{d}\mu_0(x_0) \tag{2.49}$$

for any measurable function $\phi : \Omega^2 \to [0, \infty]$ (see [76, Theorem 5.3.1]). The approximate Monge maps $\mathbf{t}_0^1$ and $\mathbf{t}_0^2$ are then used in (2.43).

When the optimal plans are not unique, one must choose among the set of optimal plans. In practice, this is determined by the algorithm used to solve the Kantorovich optimization problem Equation (2.2).

2.4.3 $\mathbf{LinW}_{2,\mathcal{R}}$ Pseudo-distance in the Discrete Setting

In practice, it's rarely possible to have a reference measure $\mu_0 \in \mathcal{M}_{1,\mathcal{L}}(\Omega)$. Instead, one usually generates something discrete, i.e., a collection of particles at locations y_i with mass R_i. Denote it as the measure $\mathcal{R} = \sum R_i \delta_{y_i}$.

Now for any input data measure $\mathcal{E}$ which is itself discrete, let r_{ij} denote an optimal transport plan from $\mathcal{R}$ to $\mathcal{E}$. Note that there may be more than one optimal transport plans between two given events. In general, a transport plan r_{ij} may send mass from particle i in the reference $\mathcal{R}$ to many different particles in $\mathcal{E}$. Consider the average of these locations, weighted by how much mass is sent to each and normalized by the amount of mass starting at particle i,

$$z_i := \frac{1}{R_i} \sum_j r_{ij} x_j \tag{2.50}$$

This provides a map from an event $\mathcal{E}$ to a vector z_i in the n-dimensional Euclidean space, $\mathbb{R}^n$, where n is the number of particles in the reference $\mathcal{R}$.

The LOT approximation of the 2-Wasserstein metric then measures the distance between two events $\mathcal{E}$ and $\tilde{\mathcal{E}}$ by considering the Euclidean distances between all pairs $(z_i, \tilde{z}_i)$, weighted by the mass starting at particle i,

$$\mathrm{LinW}_{2,\mathcal{R}}(\mathcal{E},\tilde{\mathcal{E}})|_{r,\tilde{r}} = \left(\sum_i R_i \|z_i - \tilde{z}_i\|^2\right)^{1/2}. \tag{2.51}$$

Note that this approximation explicitly depends on the choice of transport plans $r_{ij}, \tilde{r}_{ij}$ from the reference to the two measures respectively. This implies that the $\mathrm{LinW}_{2,\mathcal{R}}$ approximation does not always give the same value, due to the choice of different optimal transport plans.

As can be expected, such a $\mathrm{LinW}_{2,\mathcal{R}}$ approximation is not in general a metric on the space of measures. For instance, if the reference $\mathcal{R}$ consists of a single particle at location y_1, then $z_1 = \sum_j x_j E_j$ is the "center of mass" of $\mathcal{E}$. And any two events $\mathcal{E}, \tilde{\mathcal{E}}$ with equal center of mass satisfy $\mathrm{LinW}_{2,\mathcal{R}}(\mathcal{E},\tilde{\mathcal{E}})|_{r,\tilde{r}} = 0$. Consequently, it is clear that a necessary condition for the $\mathrm{LinW}_{2,\mathcal{R}}$ approximation to capture finer properties of the input measures is that the reference cannot be too concentrated.

In fact, this condition is also sufficient. When the reference does not concentrate on lower dimensional sets, the $\mathrm{LinW}_{2,\mathcal{R}}$ pseudo-distance obtained via the above procedure coincides with the well-defined LinW_2 metric introduced in the previous subsection. For a proof, see the Appendix of [35].

There we further proved that, if the reference $\mathcal{R}$ is given by a collection of N^2 particles, uniformly distributed on a rectangle Ω, with equally weighted masses $R_i^N = 1/N^2$, then, as $N \to +\infty$, the $\mathrm{LinW}_{2,\mathcal{R}}$ pseudo-distance converges to the true LinW_2 metric, where the reference $\mathcal{R} = \mu_0$ is now the probability measure uniformly distributed on Ω,

$$\lim_{N\to+\infty} \mathrm{LinW}_{2,\mathcal{R}}(\mathcal{E},\tilde{\mathcal{E}})|_{r^N,\tilde{r}^N} = \mathrm{LinW}_2(\mathcal{E},\tilde{\mathcal{E}}). \tag{2.52}$$

This justifies our latter use of a uniformly distributed discrete measure as the reference for our physics applications.

For such choice of $\mathcal{R} = \mu_0$ and any events $\mathcal{E}, \tilde{\mathcal{E}}$ on Ω, the transport metric is bounded above and below by the original 2-Wasserstein distance [77],

$$\mathrm{W}_2(\mathcal{E}, \tilde{\mathcal{E}}) \leq \mathrm{LinW}_2(\mathcal{E}, \tilde{\mathcal{E}}) \leq C\mathrm{W}_2(\mathcal{E}, \tilde{\mathcal{E}})^{2/15}, \tag{2.53}$$

where the constant $C > 0$ depends on Ω. In this way, $\mathrm{LinW}_{2,\mathcal{R}}$ not only converges to a well-defined transport metric LinW_2, but that transport metric also captures the behavior of the original 2-Wasserstein metric at large and small distances.

An Example Calculation of the $\mathbf{LinW}_{2,\mathcal{R}}$ Pseudo-distance

Finally let us illustrate the $\mathrm{LinW}_{2,\mathcal{R}}$ pseudo-distance and its relationship to the standard 2-Wasserstein metric with a simple example; see Figure 2.4. The two discrete measures, both consisting of two Dirac masses, are highlighted in blue and in red, respectively. The blue measure (denoted as "Jet 1" in the plot) has its two composite particles located at $(-1.0, 0.0)$ and $(-0.5, 0.0)$, whereas the particles of the red measure ("Jet 2") are at $(-1.0, 0.0)$ and $(1.0, 0.0)$. The reference measure (green; "Reference Jet") contains $9 \times 9 = 81$ particles uniformly distributed on $\Omega = [-1.0, 1.0]^2$.

The top row of Figure 2.4 shows the optimal transport plans that rearrange the Reference Jet into Jet 1 and Jet 2, respectively, according to the exact 2-Wasserstein metric (denoted as OT-W2 in the plot). Here, grey lines indicate how mass from particle y_i in the reference is sent to particle x_j in Jet 1 or particle $\tilde{x}_j$ in Jet 2. Note that, as there are multiple optimal ways to perform this rearrangement, the rearrangement is not guaranteed to be symmetric: in the top left figure, compare the fifth particle from the left on the bottom row (which splits mass between both blue particles) to the top row (which sends all mass to the right particle).

In the bottom left subplot, we illustrate $\tilde{z}_i - z_i$, to visualize the difference in how the reference is rearranged for Jet 1 and Jet 2. Predictably, we observe that the main difference is mass going further to the right in the case of Jet 2. The LOT approximation of the 2-Wasserstein distance (denoted as LOT-W2 in the plot) is computed by taking the sum of the lengths of the gray vectors squared, weighted by the mass of the reference measure $R_i = 1/81$, so that $\mathrm{LinW}_{2,\mathcal{R}}(\mathcal{E}, \tilde{\mathcal{E}})|_{r,\tilde{r}} \approx 1.07$.

Finally, in the lower right subplot, we illustrate the exact W_2 distance directly between Jet 1 and Jet 2, which corresponds to moving half of the mass in the Jet 1 a distance 1.5 as one would expect. So $\mathrm{W}_2(\mathcal{E}, \tilde{\mathcal{E}}) = (1.5^2/2)^{1/2} \approx 1.06$. It is satisfying to observe that $\mathrm{LinW}_{2,\mathcal{R}}(\mathcal{E}, \tilde{\mathcal{E}})|_{r,\tilde{r}} \approx \mathrm{W}_2(\mathcal{E}, \tilde{\mathcal{E}})$: the $\mathrm{LinW}_{2,\mathcal{R}}$ pseudodistance in this case is very close to the actual W_2 distance between the two discrete measures.

To summarize, we have learnt that the linearized W_2 metric between two measures can be obtained through the Euclidean distance between their logarithmic maps with respect to a Lebesgue-absolutely continuous measure. In actual numerical applications, one has to discretize the reference measure and the corollaries proven in the Appendix of [35] guarantees that such a $\mathrm{LinW}_{2,\mathcal{R}}$ pseudo-distance converges to the LinW_2 distance, which in turn well approximates the exact W_2 distance.

2.5 Linearized Hellinger-Kantorovich Metric

We now describe the linearization of the Hellinger-Kantorovich metric, shorthanded as “LinHK” (or sometimes “LinHK_κ” to emphasize the hyper-parameter κ). As before, we begin by exploring the Riemmanian structure of the HK distance. After defining the LinHK distance on the space of general probability measures, we focus on the special case of discrete measures, where following the notation in the case of W_2 we denote the

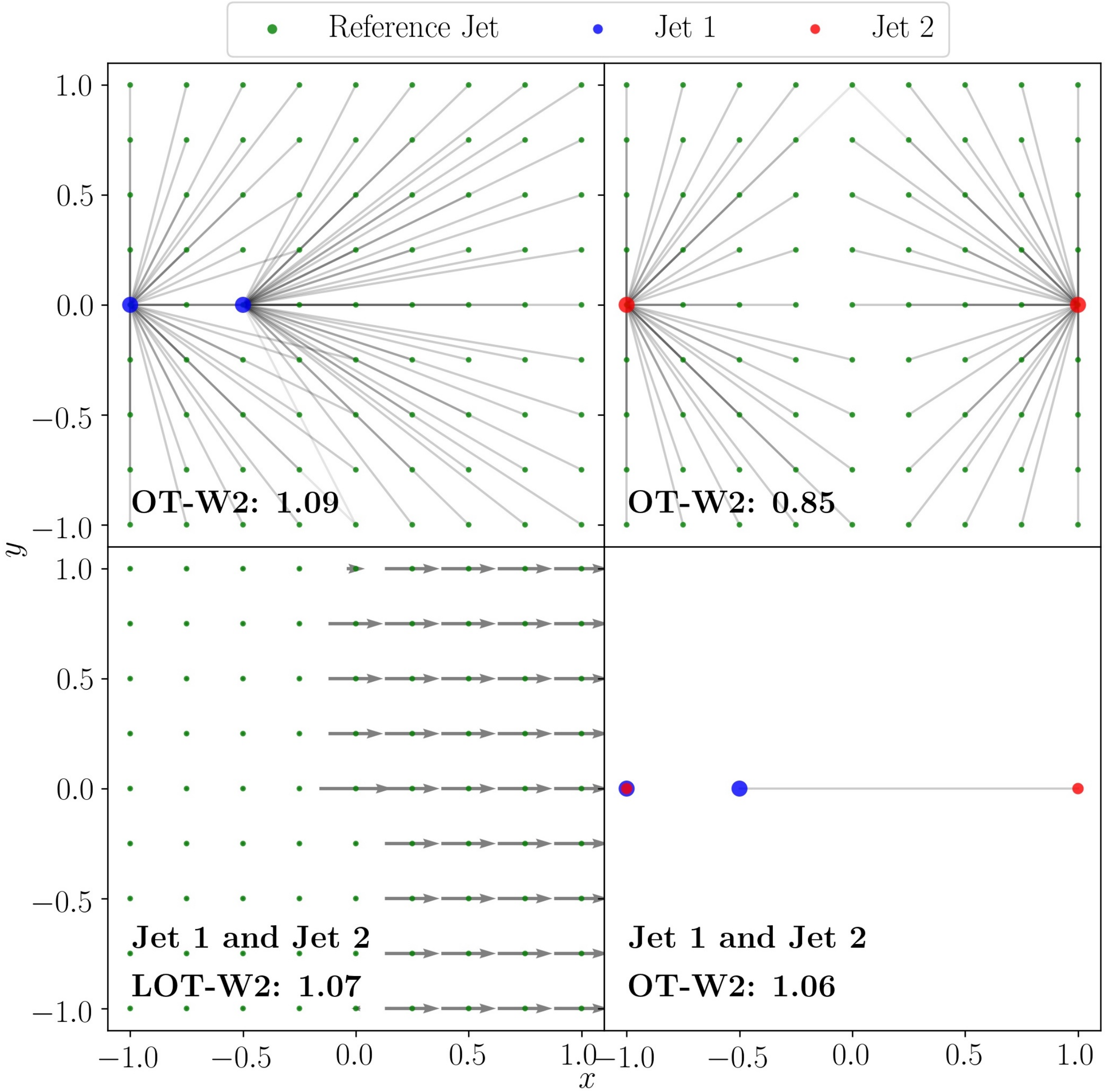


Figure 2.4: *Upper left:* An optimal movement using the exact W_2 metric to rearrange a uniform reference jet of $9 \times 9 = 81$ constituent particles (green) into the sample Jet 1 (blue). *Upper right:* An optimal movement using the exact W_2 metric to rearrange the same uniform reference jet (green) into another sample Jet 2 (red). *Lower left:* An optimal movement to rearrange the sample Jet 1 into the sample Jet 2 using LinW_2 approximation. *Lower right:* An optimal movement to rearrange the two sample jets directly using exact W_2.

corresponding LinHK approximation as $\mathrm{LinHK}_{\kappa,\mathcal{R}}$.

Most statements here are presented without proofs. Additionally, when the results themselves are too complicated and cluttered with long expressions irrelevant for our current usage, we choose to not write them down explicitly and instead refer the readers to the original paper [34] and the references therein for a full mathematical exposition.

2.5.1 Riemmanian Structure of HK

In the previous section on the 2-Wasserstein distance, we showed its Riemannian structure explicitly via Equation (2.41). Here, we seek an equivalent expression for the HK distance. In other words, we want to express $\mathrm{HK}(\mu_0, \mu_1)$ in terms of the particles' initial tangent direction at $t = 0$. Since the HK distance allows transport as well as mass changes, the tangent space will now consist of a velocity field and a mass growth field. Special care must be applied to the regions where "teleport" occurs, in particular where mass is created from nothing.

Let $\mu_0 \in \mathcal{M}_{+,\mathcal{L}}(\Omega)$, $\mu_1 \in \mathcal{M}_+(\Omega)$ be our two general probability measures. Notice the additional requirement of Lebesgue-absolutely continuousness on μ_0. By now, you should already be very familiar about the functionality of this condition.

Let π be a corresponding minimizer for $\mathrm{HK}(\mu_0, \mu_1)$ in Equation (2.27), which—thanks to the above condition on μ_0—is unique and can be written as $\pi = (\mathbf{id}, \mathbf{t})_\sharp \sigma$ for some measurable $\mathbf{t} : \Omega \to \Omega$ and $\sigma \in \mathcal{M}_+(\Omega)$. And let (ρ, ω, ζ) be the corresponding minimizers of Equation (2.24) (and let $(\tilde{\rho}, \tilde{\zeta})$ be the corresponding parts of (ρ, ζ) as given by [34, Proposition 3.14]). Then one has

$$\mathrm{HK}(\mu_0, \mu_1)^2 = \int_0^1 \int_\Omega \left[\left\| \frac{\mathrm{d}\omega_t}{\mathrm{d}\rho_t} \right\|^2 + \tfrac{1}{4} \left(\frac{\mathrm{d}\zeta_t}{\mathrm{d}\rho_t} \right)^2 \right] \mathrm{d}\rho_t \, \mathrm{d}t, \tag{2.54}$$

where we set $\kappa = 1$ for simplicity.

A subtle difference between Equation (2.54) and Equation (2.36) is that here the integrand

$$\int_\Omega \left[\left\| \frac{\mathrm{d}\omega_t}{\mathrm{d}\rho_t} \right\|^2 + \tfrac{1}{4} \left(\frac{\mathrm{d}\zeta_t}{\mathrm{d}\rho_t} \right)^2 \right] \mathrm{d}\rho_t$$

must be handled with particular care for $t \in \{0, 1\}$, as one may have that $\frac{\mathrm{d}\zeta_t}{\mathrm{d}\rho_t}$ diverges in some locations as $t \to 0$ and 1 where ρ_t vanishes in the limit $t = \{0, 1\}$. Thus, we cannot simply rewrite Equation (2.54) in terms of this integrand at $t = 0$, as we have done for W_2 in deriving Equation (2.41) from Equation (2.36).

The subtleties can be handled by Lebesgue decomposing μ_0 and μ_1 with respect to the marginals of π. That is,

$$\mu_0 = u_0 \cdot \sigma + \mu_0^\perp, \qquad \mu_1 = u_1 \cdot \mathbf{t}_\sharp \sigma + \mu_1^\perp. \tag{2.55}$$

Set further for $t \in [0, 1)$:

$$v_t := \frac{\mathrm{d}\omega_t}{\mathrm{d}\rho_t}, \qquad \alpha_t := \frac{\mathrm{d}\tilde{\zeta}_t}{\mathrm{d}\rho_t} - 2(1-t)\frac{\mathrm{d}\mu_0^\perp}{\mathrm{d}\rho_t}. \tag{2.56}$$

Then, we have

$$v_0(x) = \begin{cases} \frac{\mathbf{t}(x)-x}{\|\mathbf{t}(x)-x\|} \cdot \sqrt{\frac{u_1(\mathbf{t}(x))}{u_0(x)}} \cdot \sin(\|\mathbf{t}(x) - x\|) & \sigma\text{-a.e.,} \\ 0 & \mu_0^\perp\text{-a.e.,} \end{cases} \tag{2.57a}$$

$$\alpha_0(x) = \begin{cases} 2\left(\sqrt{\frac{u_1(\mathbf{t}(x))}{u_0(x)}} \cdot \cos(\|\mathbf{t}(x) - x\|) - 1 \right) & \sigma\text{-a.e.,} \\ -2 & \mu_0^\perp\text{-a.e..} \end{cases} \tag{2.57b}$$

with the convention that $v_0(x) = 0$ if $\mathbf{t}(x) = x$, and

$$\mathrm{HK}(\mu_0, \mu_1)^2 = \int_\Omega \left[\|v_0\|^2 + \tfrac{1}{4}(\alpha_0)^2\right] \,\mathrm{d}\mu_0 + \|\mu_1^\perp\|\,. \tag{2.58}$$

Intuitively, v_t describes the spatial movement of mass particles, α_t describes the change of mass of moving particles and of those that disappear entirely at $t = 1$. And $\mu_1^\perp$ describes the mass particles that are created from nothing.

The third singular term may be undesirable from a practical point of view. Fortunately, under certain assumptions on the relative distributions of μ_1 with respect to μ_0 (which can easily be achieved by a reasonable choice of μ_0; more to come later), we can simply drop the $\|\mu_1^\perp\|$ term as $\mu_1^\perp = 0$, and write instead

$$\mathrm{HK}(\mu_0, \mu_1)^2 = \int_\Omega \left[\|v_0\|^2 + \frac{1}{4}(\alpha_0)^2\right] \,\mathrm{d}\mu_0.$$

This holds, in particular, when for a dataset of samples $\{\mu_1, \ldots, \mu_n\}$, μ_0 is chosen as linear mean or HK-barycenter (see [78] for details).

2.5.2 LinHK Distance on the Space of Probability Measures

We now identify a candidate for the logarithmic map, given as an explicit function from an optimal Kantorovich-type transport plan. Let $\mu_0 \in \mathcal{M}_{+,\mathcal{L}}(\Omega)$, $\mu_1 \in \mathcal{M}_+(\Omega)$ and let $(v_t, \alpha_t, \mu_1^\perp)$ be given as above. We define the Logarithmic map for HK at support point μ_0 for the measure μ_1 as

$$\mathrm{Log}_{\mathrm{HK}}(\mu_0; \mu_1) := (v_0, \alpha_0, \sqrt{\mu_1^\perp}). \tag{2.59}$$

Now given another measure $\tilde{\mu}_1 \in \mathcal{M}_+(\Omega)$ with $\mathrm{Log}_{\mathrm{HK}}(\mu_0; \tilde{\mu}_1) := (\tilde{v}_0, \tilde{\alpha}_0, \sqrt{\tilde{\mu}_1^\perp})$ (we use tilde "$\tilde{\mu}_1$" instead of subscript "μ_2" to avoid potential conflict with, for example, v_0), we define the corresponding inner product as

$$g_{\mathrm{HK}}\big(\mu_0; (v_0, \alpha_0, \sqrt{\mu_1^\perp}), (\tilde{v}_0, \tilde{\alpha}_0, \sqrt{\tilde{\mu}_1^\perp})\big) := \int_\Omega \left[\langle v_0, \tilde{v}_0\rangle + \tfrac{1}{4}\alpha_0\,\tilde{\alpha}_0\right]\,\mathrm{d}\mu_0 + \int_\Omega \sqrt{\frac{\mathrm{d}\mu_1^\perp}{\mathrm{d}\lambda}\,\frac{\mathrm{d}\tilde{\mu}_1^\perp}{\mathrm{d}\lambda}}\,\mathrm{d}\lambda \tag{2.60}$$

where λ is some measure in $\mathcal{M}_+(\Omega)$ with $\mu_1^\perp, \tilde{\mu}_1^\perp \ll \lambda$.

Uniqueness of $(v_0, \alpha_0, \sqrt{\mu_1^\perp})$ is implied by the uniqueness of the optimal coupling π from which they are constructed. Hence, $\mathrm{Log}_{\mathrm{HK}}(\mu_0; \cdot)$ is well-defined. Of course, referring to Equation (2.59) and Equation (2.60) as logarithmic map and inner product is a slight abuse of notation, since the third component of the inner product is merely defined on the cone of non-negative measures and thus lacks the full vector space structure.

Now, analogous to Equation (2.42), we have

$$\mathrm{HK}(\mu_0, \mu_1)^2 = g_{\mathrm{HK}}(\mu_0; \mathrm{Log}_{\mathrm{HK}}(\mu_0; \mu_1), \mathrm{Log}_{\mathrm{HK}}(\mu_0; \mu_1)). \tag{2.61}$$

And so, in analogy to Equation (2.43), we use this to linearize HK around the support point μ_0:

$$\mathrm{LinHK}(\mu_0; \mu_1, \mu_2)^2 := g_{\mathrm{HK}}\left(\mu_0; \mathrm{Log}_{\mathrm{HK}}(\mu_0; \mu_1) - \mathrm{Log}_{\mathrm{HK}}(\mu_0; \mu_2), \mathrm{Log}_{\mathrm{HK}}(\mu_0; \mu_1) - \mathrm{Log}_{\mathrm{HK}}(\mu_0; \mu_2)\right). \tag{2.62}$$

As in the 2-Wasserstein case, we can view the linear HK distance as a distance between the formal logarithmic maps in an (almost) Euclidean space. Indeed,

$$\begin{aligned}&\mathrm{LinHK}(\mu_0;\mu_1,\mu_2)^2 = \|\mathrm{Log}_{\mathrm{HK}}(\mu_0;\mu_1)_1 - \mathrm{Log}_{\mathrm{HK}}(\mu_0;\mu_2)_1\|^2_{\ell^2(\mu_0)}\\ &+\frac{1}{4}\|\mathrm{Log}_{\mathrm{HK}}(\mu_0;\mu_1)_2 - \mathrm{Log}_{\mathrm{HK}}(\mu_0;\mu_2)_2\|^2_{\ell 2(\mu_0)} + \|\mathrm{Log}_{\mathrm{HK}}(\mu_0;\mu_1)_3 - \mathrm{Log}_{\mathrm{HK}}(\mu_0;\mu_2)_3\|^2_{\mathrm{Hell}}\,,\end{aligned} \tag{2.63}$$

where by $\|\cdot\|_{\mathrm{Hell}}$ we denote the Hellinger distance over measure square roots. And thus the linear HK distance can be embedded in the space $\ell^2(\mu_0;\mathbb{R}^d)\times\ell^2(\mu_0;\mathbb{R})\times\sqrt{\mathcal{M}_+(\Omega)}$ where the third component is the cone of square roots of non-negative measures, equipped with the Hellinger metric.

As hoped, when μ_0 has sufficiently wide support on Ω, the third component is always zero and the embedding can be made into the Euclidean space $\ell^2(\mu_0;\mathbb{R}^d)\times\ell^2(\mu_0;\mathbb{R})$ where $g_{\mathrm{HK}}(\mu_0;\cdot,\cdot)$ is an inner product. Still, a more careful look at this singular measure-valued component is needed in the future to better understand its behavior.

From a practical perspective, the additional complexity of LinHK distance relative to the LinW$_2$ distance is relatively low. Loosely speaking, one must apply an unbalanced version of the Sinkhorn algorithm, adjust the formula for the initial velocity field, accommodate an additional scalar mass change field, and finally fix a single real-valued length-scale parameter κ by validation on the data. All these steps only introduce a small amount of computational overhead, in comparison to the exact OT calculation itself.

As the Hellinger-Kantorovich space is much less understood than the Wasserstein space, there still remain many open questions left for future mathematical study. For example, we know that the linearization in the 2-Wasserstein space is closed under convex combinations. In other words, if $\{v_i\}_{i=1}^n$ are a set of W$_2$ linear embeddings, then any $\tilde{v}$ in the convex hull of $\{v_i\}_{i=1}^n$ is in the domain of the exponential map and hence one can generate a new measure via $\tilde{\mu}=\mathrm{Exp}_{\mathrm{W}_2}(\tilde{v})$ (see Section 3.3 for applications to data augmentation for jets). Identifying similar operations under which the HK linearization is closed is definitely worth pursuing.

Another important open problem is to quantitatively bound the accuracy of the linear approximation of the HK distance, which requires an estimation of the curvature of the HK manifold. This would hopefully enable us to obtain an upper and lower bound similar to Equation (2.53).

2.5.3 LinHK$_{\kappa,\mathcal{R}}$ Pseudo-distance in the Discrete Setting

We now focus on the case of discrete measures. Let $\mathcal{R}$ be a discrete reference measure, consisting of particles at locations $\{x_i\}_{i\in I}$ with positive masses $\{R_i\}_{i\in I}$. For any discrete measure $\mathcal{E}$, let γ_{ij} denote an optimizer of Equation (2.28), which represents an optimal transport plan from the auxiliary measures $\mathcal{G}$ to $\tilde{\mathcal{G}}$. Note that more than one optimizer may exist.

In general, the transport plan γ_{ij} may send mass from x_i in $\mathcal{G}$ to many different locations in $\tilde{\mathcal{G}}$. In order to linearize the HK metric, we first consider the average of these locations, weighted by how much mass is sent to each place and normalized by the amount of mass starting at x_i in $\mathcal{G}$,

$$z_i = \begin{cases} \frac{1}{G_i}\sum_j \gamma_{ij}\tilde{x}_j & \text{if } G_i > 0 \\ x_i & \text{if } G_i = 0 \end{cases} \tag{2.64}$$

Next, we consider the average amount that mass starting at location x_i needs to be rescaled, via creation or destruction, in order for $\mathcal{R}$ to become $\tilde{\mathcal{E}}$: For each $\tilde{x}_j$, consider the ratio $\tilde{E}_j/\tilde{G}_j$, between the amount of mass that must end up at location $\tilde{x}_j$ and the amount of mass transported by γ_{ij} to $\tilde{x}_j$. If $\tilde{E}_j/\tilde{G}_j > 1$, mass needs to be created at x_j, and if $\tilde{E}_j/\tilde{G}_j < 1$, mass needs to be destroyed at x_j. Note that this quantity is well-defined only for $\tilde{G}_j = \sum_i \gamma_{ij} > 0$. In fact, this is a necessary assumption for the Hellinger-Kantorovich metric to be linearized in a manner that admits a Euclidean embedding [34, p18].

Recall that a sufficient condition for $\gamma_{ij} > 0$ is $d_{ij} = \|x_i - \tilde{x}_j\| < \frac{\kappa\pi}{2}$. Consequently, in what follows, we will suppose that κ is sufficiently large so that, for each $\tilde{x}_j$, there exists x_i so that

$$\|x_i - \tilde{x}_j\| < \frac{\kappa\pi}{2}. \tag{2.65}$$

This will ensure $\tilde{G}_j > 0$ for all j.

With this assumption in hand, we now consider, for each fixed x_i, the weighted average of this ratio, representing how much mass needs to be created/destroyed at x_j, with respect to how much mass γ_{ij} transports to each $\tilde{x}_j$, normalized by the amount of mass G_i originally starting at x_i:

$$u_i = \begin{cases} \frac{1}{G_i} \sum_j \left(\frac{\tilde{E}_j}{\tilde{G}_j}\right) \gamma_{ij} & \text{if } G_i > 0, \\ 0 & \text{if } G_i = 0. \end{cases} \tag{2.66}$$

Intuitively, while the coordinate z_i, defined in Equation (2.64), represents the average location that mass starting at x_i is transported to in $\tilde{E}$, the coordinate u_i represents the average amount of creation/destruction that will happen to mass that started at x_i, after it is transported.

With these quantities in hand, we may now state the formula for the linearized Hellinger-Kantorovich approximation with respect to a discrete measure. In the original paper [36], LinHK in the discrete case is given the acronym pluOT (particle linearized unbalanced Optimal Transport), in order to emphasize that it is a discrete particle approximation of the continuum linearization of the Hellinger-Kantorovich metric. Here, to avoid overuse of terminology, we refrain from introducing pluOT and instead keep the notation $\mathrm{LinHK}_{\kappa,\mathcal{R}}$ to maintain the parallelism with the $\mathrm{LinW}_{2,\mathcal{R}}$ case.

We now define the $\mathrm{LinHK}_{\kappa,\mathcal{R}}$ distance as

$$\begin{aligned} &\mathrm{LinHK}_{\kappa,\mathcal{R}}(\mathcal{E},\tilde{\mathcal{E}}) \qquad (2.67)\\ &\quad= \left(\sum_i R_i\|v_i-\tilde{v}_i\|^2+\frac{\kappa^2}{4}R_i|\alpha_i-\tilde{\alpha}_i|^2\right)^{1/2},\\ &v_i=\kappa\ \mathrm{sgn}(z_i-x_i)\sqrt{u_iG_i/R_i}\sin(\|z_i-x_i\|/\kappa),\\ &\alpha_i=2\left(\sqrt{u_iG_i/R_i}\cos(\|z_i-x_i\|/\kappa)-1\right). \end{aligned}$$

Note that this approximation depends on the choice of the optimal transport plans $\gamma_{ij},\tilde{\gamma}_{ij}$ via their dependence on $x_i, z_i, \tilde{x}_i, \tilde{z}_i$; see Equations (2.64) and (2.66).

As in the definition of HK, the unusual expressions for v_i and α_i in $\mathrm{LinHK}_{\kappa,\mathcal{R}}$ derive from the surprising connection to cone geometry [51, 52, 53]. In particular, when comparing the locations and masses of particles (x_i, E_i), the cone structure is used to identify all points with mass zero as the same point. For example, in one spatial dimension (and under assumption (2.65)), (x_i, E_i) corresponds to the point $(E_i\cos(x_i/\kappa), E_i\sin(x_i/\kappa))$ in the plane.

To see the connection with Equation (2.67), consider the original location and mass of the ith particle, (x_i, R_i), along with the average location to which its mass is sent and the average mass at that location after creation/destruction, (z_i, u_iG_i). The *constant speed geodesic* in the *cone metric* between these two points represents how the location x_i is optimally transported to z_i, while simultaneously mass is created and destroyed to convert R_i into u_iG_i [52]. In one spatial dimension, this is just the line connecting the two points in the plane.

The coordinate v_i represents the velocity of the spatial trajectory at time zero, while $R_i\alpha_i$ represents the rate of change of the mass at time zero. From this perspective, $\mathrm{LinHK}_{\kappa,\mathcal{R}}$ measures the difference between two events $\mathcal{E}$ and $\tilde{\mathcal{E}}$ in terms of how a reference event $\mathcal{R}$ deforms into $\mathcal{E}$ and $\tilde{\mathcal{E}}$, by comparing the velocities by which particles in the reference event move and the rates at which their masses change.

In analogy with $\mathrm{LinW}_{2,\mathcal{R}}$, a key benefit of the linear approximation of the Hellinger-Kantorovich metric is that it provides a natural embedding

$$\mathcal{E} \mapsto (v_i, \alpha_i)_{i\in I} \in \mathbb{R}^{dn} \times \mathbb{R}^n, \tag{2.68}$$

where d is the dimension of the underlying domain Ω in which particles are located and n is the number of particles in the discrete reference measure, $n = |I|$. This vector may be interpreted geometrically as an approximation of the tangent vector from $\mathcal{R}$ to $\mathcal{E}$ with respect to the Hellinger-Kantorovich geometry, an interpretation that may be made precise when $\mathcal{R}$ is a finite Borel measure that is absolutely continuous with respect to Lebesgue mesaure, as explained above [34, Definition 4.5]. In this way, it is natural to compare two discrete measures $\mathcal{E}$ and $\tilde{\mathcal{E}}$ by computing the distance between the vectors (v_i, α_i) and $(\tilde{v}_i, \tilde{\alpha}_i)$ as elements of the tangent space at $\mathcal{R}$, as in Equation (2.67) above.

When later we use $\mathrm{LinHK}_{\kappa,\mathcal{R}}$ as a tool for data analysis, we will investigate the effects of creation/destruction in the HK metric separately from the fact that it allows for the comparison of measures with unequal total masses. We do this by separately analyzing the statistical performance of the linearization of $\mathrm{LinHK}_{\kappa,\mathcal{R}}(\mathcal{E}, \tilde{\mathcal{E}})$ with the performance of $\mathrm{LinHK}_{\kappa,\mathcal{R}}\left(\mathcal{E}/\sum_i E_i, \tilde{\mathcal{E}}/\sum_j \tilde{E}_j\right)$, where $\mathcal{E}/(\sum_i E_i)$ denotes the normalized measure, in which the mass E_i of each particle in $\mathcal{E}$ is replaced by $E_i/(\sum_i E_i)$.

The HK metric exhibits a simple scaling under the above transformation [63, Theorem

3.3]: Denoting $m = \sum_i E_i$,

$$\begin{aligned} (\tilde{E}_j)^{\mathrm{norm}} &= m^{-1/2}\tilde{E}_j, & \gamma_{ij}^{\mathrm{norm}} &= m^{-1/2}\gamma_{ij}, \\ G_i^{\mathrm{norm}} &= m^{-1/2}G_i, & (\tilde{G}_j)^{\mathrm{norm}} &= m^{-1/2}\tilde{G}_j, \\ z_i^{\mathrm{norm}} &= z_i, & u_i^{\mathrm{norm}} &= m^{-1/2}u_i, \\ v_i^{\mathrm{norm}} &= m^{-1/2}v_i, & \alpha_i^{\mathrm{norm}} &= m^{-1/2}\alpha_i + 2(m^{-1/2} - 1). \end{aligned} \tag{2.69}$$

For future study, one would like to better understand the stability of the Hellinger-Kantorovich maps with respect to discretization and therefore the stability of the linear HK distance with respect to discretization. The corresponding stability study for the 2-Wasserstein distance was established in [79].

2.6 Optimal Transport in Action: A Numerical Example

To help digest the rich mathematics developed in this chapter, here we present a numerical example on synthetic images as an illustration. First, we apply both linearized W_2 and linearized HK embeddings on ellipses with various sizes and elongations. We then analyze the resulting LOT manifolds using a simple Principal Component Analysis (PCA), which clearly demonstrates the advantage of the $\mathrm{LinHK}_{\kappa,\mathcal{R}}$ over the $\mathrm{LinW}_{2,\mathcal{R}}$ metric, at least for the present example.

This section serves as a prelude to the following two chapters, where LOT embeddings will be applied to much more complicated datasets in collider physics and astrophysics. The statistical analysis frameworks will also be more advanced and tailored for the particular physics use case. Hopefully, the current example of ellipses can illuminate some key aspects of the linearization of the two OT metrics, transforming the mathematical jargons throughout the chapter into useful intuitive understanding for practical applications.

2.6.1 Synthetic Data: Deforming and Resizing Ellipses

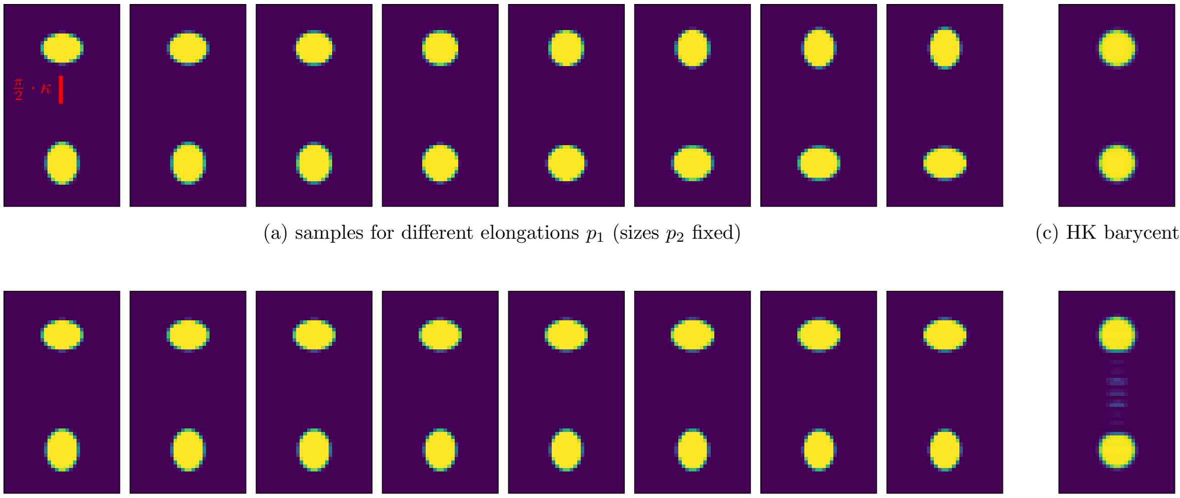
We generate a synthetic dataset with each set of samples consisting of the images of two ellipses on a 64×64 pixel grid. The mass density is set to be 1 within the ellipses and zero outside. At the boundaries, the density is non-binary and between 0 and 1 due to rasterization effects. Ellipses are first rendered at a higher resolution and then reduced to the standard 64×64 pixels.

Each image of ellipse is characterized by two parameters $p_1, p_2 \in [-1, 1]^2$. Here p_1 specifies the elongation of the ellipses. When $p_1 = 0$, both ellipses are reduced to circles. For $p_1 > 0$, one becomes elongated horizontally, the other one vertically. For $p_1 < 0$, the roles are reversed.

The other parameter p_2 controls the resizing of the two ellipses, and therefore their relative masses. When $p_2 = 0$, the sizes of two ellipses are equal. For $p_2 > 0$, one ellipse expands and the other shrinks, whereas the role is again reversed for $p_2 < 0$. The maximal change in the ellipse diameter between $p_2 = -1$ and $p_2 = +1$ is approximately 0.5 pixels, with the corresponding relative change in mass being approximately 10%.

Examples for different pairs of (p_1, p_2) are shown in Figure 2.5 (a,b). They are generated by sampling both parameters on 8 equidistant points from $[-1, 1]$, yielding a total of $n = 64$ input images. The resizing in Figure 2.5 (b) is not very noticeable, due to the minute pixel variation (0.5 as quoted above). Note that prior to any analysis, all images are normalized to have a total mass of one.

We can regard each sample as a sum of Dirac measure at pixel locations, i.e., $(\mathcal{E}_i)_{i=1}^n$ with n being the total number of samples. Obviously, each $\mathcal{E}_i$ is a measure on the image



(a) samples for different elongations p_1 (sizes p_2 fixed)

(c) HK barycenter

(b) samples for different sizes p_2 (elongations p_1 fixed)

(d) W_2 barycenter

Figure 2.5: Some representative samples from the ellipses dataset and the corresponding barycenters in the HK and W_2 distances. The red line in the top left sample has length $\frac{\pi}{2} \cdot \kappa$, representing the maximal transport distance under HK. Figure copied from [34].

domain Ω, which is a rectangle in $\mathbb{R}^2$.

2.6.2 The LOT Embeddings

To apply the LOT formalism, we first need to choose a reference measure $\mathcal{R}$, against which the exact W_2 or HK distances are computed for all the images. This choice is important for a successful analysis. As both metric spaces are (weakly) curved Riemannian manifold, we expect the local linearization to be a poor approximation if the reference point $\mathcal{R}$ is very far from the samples $(\mathcal{E}_i)_{i=1}^n$. Similarly, another failure mode occurs when the samples themselves are too far from each other.

A natural candidate for $\mathcal{R}$ is the barycenter of the samples. The W_2 barycenter was analyzed in [64] and here we use the algorithm described in [80]. The corresponding HK barycenter was recently studied in [81, 78], which, in principle, can be approximated numerically with the methods from [54, 82]. However, for large numbers of samples n or when samples live on large grids, computing the barycenter could be numerically prohibitive. Thus it is worthwhile to consider alternative choices.

In [70], it was proposed to use the linear average of the samples, and one could similarly consider the Hellinger mean. In [35], a uniform measure on a Cartesian background-grid was used. The later two will be employed in the more complicated physics applications considered in the following chapters. In all three cases, it is ensured that $\mathcal{E}_i^\perp = 0$ so that we can safely ignore this singular component for the HK distance.

When applying the exponential map at a discrete reference measure $\mathcal{R}$, the result will be a discrete measure where the points of $\mathcal{R}$ are moved to new locations with their masses re-scaled. Often, it is desirable to visualize the new measure on a fixed reference grid, e.g., the same grid that the input samples live on. For rasterization, we use bilinear interpolation coefficients to distribute mass to nearby grid points.

We now analyze this dataset with $\mathrm{LinW}_{2,\mathcal{R}}$ and $\mathrm{LinHK}_{\kappa,\mathcal{R}}$. For HK, we set the length-scale parameter κ to 5, so that the maximal transport distance is sufficient to track the deformation of the ellipses induced by p_1 while still separating well the two ellipses from each other (cf. Figure 2.5). In more realistic applications such as those in the later chapters, a range of values for the κ parameter are tested over several orders of magnitude to pick the optimal κ, which is then validated to give stable performance in its vicinity.

Given the simplicity of our current example, we can afford calculating the (approximate) W_2 or HK barycenter of the samples. We thus use them as the reference point $\mathcal{R}$, visualized in Figure 2.5 (c,d). Note that due to the difference of the masses of the two ellipses caused by p_2 variation, the W_2 barycenter has some mass located in between the two disks, where the discrete pattern stems from the fact that we used only a finite number of values for p_2. The HK barycenter exhibits no such artifacts and is thus more desirable.

Once the reference point is fixed, we obtain the linear embeddings via the logarithmic maps, where we set $w_{i,0} := \mathrm{Log}_{\mathrm{W}_2}(\mathcal{R}; \mathcal{E}_i)$ and $(v_{i,0}, \alpha_{i,0}) := \mathrm{Log}_{\mathrm{HK}}(\mathcal{R}; \mathcal{E}_i)$ for $i = 1, \dots, n$.

2.6.3 PCA on the LOT Manifolds

We now apply a simple Principal Component Analysis (PCA) to the LOT manifolds resulting from both linear embeddings. Intuitively, if we pick the barycenter as support point for the linearization, the embeddings should already be centered in the tangent space. For the W_2 metric, this is known to be true [64, Equation (3.10)]. For the HK metric, we are not aware of such a result, but numerically it seems to be satisfied. If any other reference measure is picked, we then need to center the samples before applying PCA.

The coordinates of the (centered) linear embeddings with respect to the two dominant PCA modes are shown in Figure 2.6. For the HK metric, we recover a two-dimensional grid structure that corresponds precisely to the two underlying parameters of the dataset. In addition, the first two PCA modes capture 95% of the dataset variance.

On the other hand, for the W_2 metric, the first two principal modes only explain 78% of the total variance. The coordinates with respect to the first two principal components are dominated by the size variation. The samples lie approximately on a one-dimensional curve, according to their size variation parameter p_2. The elongation variations only cause small perturbations near this curve. Extracting information about the elongation will thus be decidedly more difficult from the W_2 embedding, in comparison to the HK embedding.

Furthermore, in Figure 2.7 the first two principal components for both linearizations are visualized as (colored) quiver plots and as curves of measures generated via the exponential map. In agreement with the previous observations, for W_2 the first two modes seem to be concerned mostly with moving mass between the two disks. For HK,

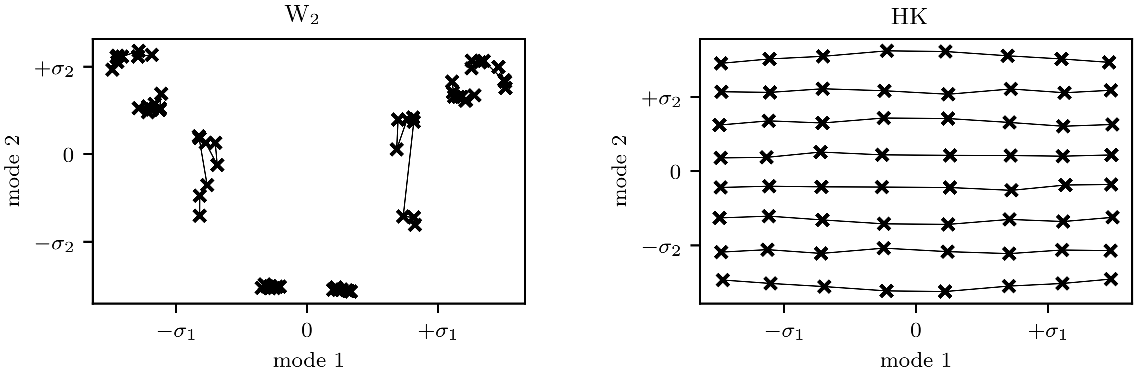


Figure 2.6: Coordinates of ellipse samples in PCA basis (two most dominant modes, axis scaling given in terms of standard deviation along each mode) for linearized W_2 and HK embeddings. Samples with identical size parameter p_2 are connected by lines. For HK, a two-dimensional grid structure emerges, one mode corresponding to elongation and the other to size. For W_2, the size variation dominates the embedding. Samples are roughly located on a one-dimensional curve, according to their size variation parameter p_2. Along this line, the samples are grouped into small clusters, each corresponding to one 'pass' through the elongation parameter p_1 for fixed p_2. Figure copied from [34].

the first mode clearly encodes variations in the disks elongation and the second mode encodes variations in their size.

Now consider picking a different reference measure, either being the ℓ^2 or Hellinger mean of the samples, or a uniform measure on Ω. For the linearized W_2 embedding, the picture is qualitatively very similar: the size variation shadows the elongation variation. For linearized HK, the situation remains essentially unchanged with ℓ^2 or Hellinger mean. For the uniform measure, the mass variation is too subtle and is shadowed by the elongation variation. However, with larger size variations, a similar picture as above re-emerges.

Of course, for more realistic datasets it cannot be expected that PCA will yield as transparent and simple results as for this toy example. We will therefore consider more sophisticated statistical models later. But the present simple example already shows us the clear advantage of HK relative to W_2. We may still hope that the ability to locally

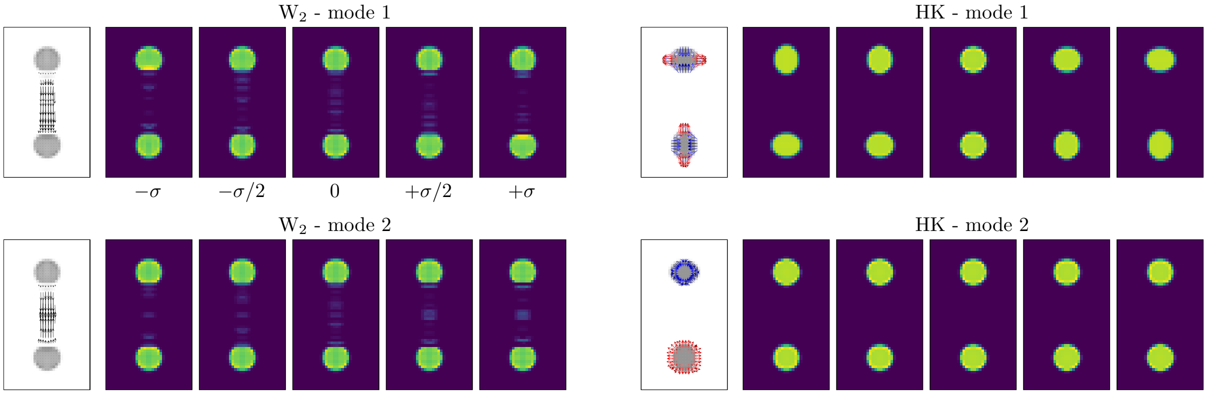


Figure 2.7: Visualizations of the two dominant PCA modes for linearized W_2 and HK embeddings of the ellipse dataset. For each mode, the quiver plot on the left shows the initial velocity field v_0, for HK the color of the arrows encodes α_0 (blue means decrease, red increase of mass). The five images on the right visualize the exponential map evaluated between $-\sigma$ and σ where σ denotes the standard deviation along the considered mode. HK accurately captures the two-dimensional structure of the dataset, W_2 is dominated by the size variations. Figure copied from [34].

vary masses will make the linear embeddings via $\mathrm{Log}_{\mathrm{HK}}$ more robust to mass fluctuations and consequently simplify any subsequent analysis task, such as classification.

Chapter 3

Optimal Transport for Collider Physics

The first full-scale application of optimal transport in high energy physics focuses on collider phenomenology. The present chapter summarizes a series of projects the author has been working on—in particular, the three publications [35, 34, 36] and some current work-in-progress. After reviewing the necessary collider physics background in Section 3.1, the following Section 3.2 develops our physics-inspired framework of machine learning with linearized optimal transport, building up the entire analysis pipeline for later statistical tasks such as jet tagging.

The next three sections 3.3, 3.4, 3.5 consist of the main body of our results, where we examine the performance of different optimal transport distances on a variety of jet tagging tasks. Section 3.6 studies the effect of pileup on optimal transport based framework. The presentation here follows closely that of our own papers, with a slight change of notations in order to maintain consistency within the thesis.

Finally, Section 3.7 concludes the chapter with a brief summary of our ongoing efforts to augment the OT framework, with the ultimate goal being full event-level classification.

It serves as a short-term roadmap for the author's projects along this line of research.

3.1 Jet Physics at the Large Hadron Collider

The Large Hadron Collider (LHC) at CERN currently provides the most stringent test on the Standard Model (SM) of particle physics at the TeV scale and represents our best hope in search of new physics beyond the Standard Model (BSM). Being a proton-proton-collision machine, the LHC at its heart produces copious high-energy quarks and gluons. Though themselves unobservable, these partons subsequently fragment into final-state hadrons recorded by the detectors and giving the experimental outputs. Hadrons deriving from the same mother parton tend to travel along the same direction within a narrow cone, resulting in a collimated spray of particles. This is called a jet, an easily identifiable structure when looking at an event display.

Over the last decades, the study of jets has emerged as a vital aspect of the LHC research program. The rich substructure of a jet becomes a window through which one can hopefully take a glimpse at the hard collisions otherwise inaccessible to observation. A plethora of jet substructure observables have since been handcrafted to probe different aspects of the underlying physics. For example, some observables are particularly sensitive to the number of prongs within a single jet and therefore serve as an excellent tool to distinguish boosted heavy particles from QCD background. Such theoretically motivated observables calculable from first-principles remain essential to deepen our understanding of the physics processes at the hadron collider and has a broader impact on QCD and beyond.

With the recent advent of machine learning, we have at out disposal a new set of powerful tools specialized at analyzing data with rich and complex patterns. As jets are naturally amenable to ML analyses, it is no surprise that this field has pioneered the adoption of ML techniques in HEP. Jet physics has now become the archetypical playground where novel statistical frameworks are developed and tested.

3.1.1 Jet Definitions

Before any analysis, a precise definition of a jet is first required in order to utilize it as a quantitative tool for collider physics. Clearly, a visual identification is far from enough. One needs a set of well-defined rules to map experimental observables—be it final-state particles or calorimeter towers—to a jet, such that its behavior is understood and reproducible. Such a set of rules is called a *jet definition*, or a *jet algorithm*. In its essence, a jet definition is nothing but a clustering algorithm, which is an example of unsupervised machine learning. In some sense, a jet itself is only defined in the context of machine learning, though the requirement that a jet needs to be physically meaningful puts stringent constraints on the validity of proposed clustering algorithms.

Usually, a jet definition involves two steps. First, it decides which particles are to be grouped together depending on some free parameters set by the user. It then assigns a momentum and/or other properties to the resulting jet. The latter step is called a recombination scheme. Various jet definitions differ in their particular choices of the particle grouping scheme and the recombination scheme. In general, they fall into one of the two general categories—cone algorithms and sequential recombination algorithms.

The very first jet finding algorithm, proposed by Sterman and Weinberg in 1977 for e^+e^- collisions [83], is a cone algorithm. Essentially, it classifies a collision event as, for example, containing two jets if all but a fraction ϵ of the total energy of the event falls within two cones of half angle δ. Here, we have two free parameters ϵ and δ, where ϵ gets rid of radiation noises and δ indicates how close two particles must be for them to be in the same jet. This top-down approach matches well with our intuition and is based on the idea that hadronization leaves the bulk features of a parton's energy flow essentially unchanged.

Most modern jet algorithms belong to the sequential recombination class. Starting from bottom-up, these algorithms compute the distance between a pair of particles, merge the closest ones, and repeat the procedure until a pre-defined stopping criterion is met. Central to such algorithms is the distance measure between particles, which needs to be physically motivated and reflect the structure of divergences in perturbative QCD. Here we give three widely used examples of sequential recombination algorithms.

First, the k_T algorithm defines its distance measure between a pair of particles i and j as

$$d_{ij} = d_{ji} = \min(p_{Ti}^2, p_{Tj}^2)\frac{\Delta R_{ij}^2}{R^2}, \tag{3.1}$$

with p_{Ti} being the transverse momentum of particle i with respect to the beam direction (the z axis) and ΔR_{ij} being the Euclidean distance on the pseudo-rapidity and azimuthal angle $(y-\phi)$ plane, i.e., $\Delta R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$. Later, we will see that this physical $y-\phi$ plane is also a suitable ground space in the definition of optimal transport distances.

Here the free parameter of the algorithm is the jet radius R which sets the angular reach of the jet. Usually, R is chosen to be 0.4 or 0.6 for small-radius jets, and 1.0 for fat jets if one would like to include as many decay products of a boosted heavy particle as possible.

The k_T algorithm also defines a similar distance between every particle i and the beam z-axis via

$$d_{iB} = p_{Ti}^2. \tag{3.2}$$

In the case where d_{ij} is smaller than d_{iB}, we merge the two particles i and j into one object called a "pseudojet", where its momentum is simply defined to be $p_i + p_j$. This so-called E-scheme is the default recombination scheme implemented in FastJet [84] and is also currently used at the LHC. Other recombination schemes also exist and are to some extend weighted versions of the E-scheme.

On the other hand, when d_{iB} is the smaller one, we simply declare particle i to be a final jet and remove it from the list. The whole procedure is repeated until a stopping criteria is reached. For the exclusive k_T algorithm, it stops when the smallest of d_{ij} or d_{iB} is above the threshold d_{cut} and all pseudojets left are then declared to be the jets of the event. For the inclusive version, no d_{cut} is necessary and the procedure terminates naturally when there are no more particles. Then among the final jets, only those with transverse momentum above a certain value are retained as the event's jets.

The anti-k_T algorithm works very similar as the k_T algorithm, except now the distance measures are

$$\begin{aligned} d_{ij} &= \min(1/p_{Ti}^2, 1/p_{Tj}^2)\frac{\Delta R_{ij}^2}{R^2}, \\ d_{iB} &= 1/p_{Ti}^2. \end{aligned} \tag{3.3}$$

Defined through $1/p_T^2$, the anti-k_T algorithm begins clustering from the hardest particles and thus has the desirable property that its hard jets are precisely circular on the $y - \phi$ plane. This is the default jet algorithm (with the E-scheme) used at all LHC experiments and is also our choice in the following study.

Finally, the Cambridge/Aachen (C/A) algorithm is even simpler, as the distances are now defined without reference to the particle p_T, i.e.,

$$\begin{aligned} d_{ij} &= \frac{\Delta R_{ij}^2}{R^2}, \\ d_{iB} &= 1. \end{aligned} \tag{3.4}$$

The above three algorithms can be unified via

$$\begin{aligned} d_{ij} &= \min(p_{Ti}^{2p}, p_{Tj}^{2p})\frac{\Delta R_{ij}^2}{R^2}, \\ d_{iB} &= p_{Ti}^{2p}, \end{aligned} \tag{3.5}$$

where $p = 1$ gives the k_T algorithm, $p = 0$ gives the C/A algorithm, and $p = -1$ gives the anti-k_T algorithm. Other p values output generalized k_T algorithms.

3.1.2 Jet Substructure Observables

Jet substructure utilizes the internal radiation pattern of a jet in order to better understand its partonic origin. The distribution of energy is referred to as the *energy flow* defined as

$$\mathcal{E} = \sum_{i=1}^{N} E_i \delta(\hat{n} - \hat{n}_i), \tag{3.6}$$

where E_i is the energy of the ith particle in a jet with N particles and $\hat{n}_i$ is its angular direction. In the context of a hadron collider, E_i is usually replaced by the transverse momentum p_{Ti} of the particle. Note that the energy flow only utilizes the experimentally measurable kinematic information and is not sensitive to, for example, particle charge and flavor.

Oftentimes, one is interested in distinguishing jets coming from different underlying hard processes, for example, whether a jet has a QCD origin or is the decay products of a boosted massive particles. This task is termed as jet tagging and is key to our understanding of the collider processes. A good jet substructure observable therefore must be able to capture the subtle difference of the radiation pattern between signals and backgrounds and should be amenable to first-principle calculations. The theory group has long paid particular attention to a property called infrared and collinear (IRC) safety, which guarantees calculability in the perturbation theory of QCD using both fixed-order calculations and resummation.

Intuitively speaking, IRC safety requires an observable $\mathcal{O}$ to be unchanged after the addition of an arbitrary number of infinitely soft partons (infrared safety) or/and an arbitrary number of collinear splittings (collinear safety). A common statement of (near) IRC safety reads

$$\mathcal{O}(p_1, p_2, ..., p_N) = \lim_{\epsilon \to 0} \mathcal{O}(\epsilon p_0, p_1, p_2, ..., p_N), \tag{3.7}$$

$$\mathcal{O}(p_1, p_2, ..., p_N) = \lim_{p_0 \to p_1} \mathcal{O}(\lambda p_0, (1-\lambda) p_1, p_2, ..., p_N), \tag{3.8}$$

where p_i is a particle's four-momentum, ϵ is a number close to 0, and $\lambda \in [0, 1]$ is the collinear splitting fraction. Although a central notion in collider physics, this definition of IRC safety—and similarly many other attempts of a mathematically rigorous formulations—suffer from various pathologies such as the inability to include multiple soft or collinear splittings. It is conjectured that optimal transport may provide a new geometric definition of IRC safety free from such pathologies; see [13] for more details.

There are many existing IRC safe observables both at the full event-level and at the jet substructure level. Here we focus on one specific well-known jet substructure observable which is particularly pertinent to our later discussions—N-subjettiness τ_N [85]. Roughly speaking, N-subjettiness counts how many subjets a given jet contains. The intuition behind it is that a QCD jet (oftentimes serving as the background) usually displays a one-prong structure, while the decay products of boosted W, Z, and Higgs bosons have two prongs and a jet resulting from the decay of a boosted top quark is more likely to have three prongs.

Formally, N-subjettiness for a jet with M particles is defined as

$$\tau_N^\beta = \min_{\hat{n}_1,...\hat{n}_N} \sum_{i=1}^{M} E_i \min(\theta_{i1}^\beta, \theta_{i2}^\beta, ..., \theta_{iN}^\beta), \tag{3.9}$$

where again E_i is replaced by p_{Ti} for a hadron collider. Here, θ_{i1} to θ_{iN} are the angular distances defined between the ith particle and the proposed subjet axes $\hat{n}_1$ through $\hat{n}_N$, which in the usual case is the Euclidean distance on the $y - \phi$ plane, i.e., $\theta_{i1} = \Delta R_{i1} = \sqrt{(\Delta y)^2 + (\Delta\phi)^2}$. The weight β is usually set to 1, but can also be changed.

The inner minimization in Equation (3.9) penalizes particles far away from any proposed axis, whereas the outer minimization searches for the optimal location of the N axes. If a jet has $\tau_N \approx 0$, then its particles are closely aligned with the candidate subjets, which means that it has N or fewer subjets. On the other hand, a $\tau_N \gg 0$ indicates that the jet has more of its energy away from the subjets and thus should have at least $N + 1$ subjets.

In practice, it is often the ratio τ_N/τ_{N-1} that is most effective at identifying N-pronged jets. For example, τ_2/τ_1 is very successful at discriminating two-pronged objects against one-pronged QCD background. Therefore, we usually use τ_N/τ_{N-1} as a benchmark to gauge the performance of our newly proposed framework. In [13], a reformulation of N-subjettiness is given based on the geometric language of optimal transport, as well as for a variety of other substructure observables.

3.1.3 Jet Tagging with Machine Learning

Machine learning provides a data-driven alternative approach which complements the above theory-centered, first-principles understanding of jet substructure. In the language of ML, jet tagging is a typical binary classification task, where a plethora of models exist using low-level inputs. For example, one can represent a jet naturally as an image using the raw data of detector hits with finite cell granularity, in which case the most suitable architecture is a convolutional neural network. This is indeed the first neural network application in jet physics [86].

Other NN architectures that have been tried over the years include recurrent neural network for jets represented as sequences, graph neural networks for jets represented as graphs, and deep sets for jets represented as point clouds such as the Energy Flow Network (EFN) and Particle Flow Network (PFN) [87]. Please refer to [88] for an overview and [89] for a more detailed exposition.

As can be seen, the central question here is how to represent jets in a suitable way that retains as much information as possible. The traditional jet substructure observables can be seen as one (or low) dimensional representations of a jet based on physical considerations. The usage of neural networks allows one to explore more complex hidden structures inside a jet and therefore unsurprisingly offers a significant improvement on classification performance.

However, deep neural networks usually function as a black-box and it is hard to understand why they work so well, which hinders further performance improvement and the ultimate extraction of physics insights. This concern has led people to design more physics-inspired learning models which may enjoy both the high performance of NNs and the theoretical interpretability of traditional observables.

One particular example is the Energy Flow Polynomials (EFP) introduced in [90].

EFPs are a complete set of observables that can be proved to linearly span the space of IRC-safe observables. As such, they can serve as inputs to simple linear regression models to learn the best set of basis functions for a particular tagging task. Physics-inspired representations like the EFPs can vastly simplify the subsequent machine learning model, while still achieving surprisingly good performance on the same par with NNs.

It is with this spirit that we introduce optimal transport for the analysis of collider events. In the following, we will see how equipping the space of jets with a well-defined notion of distance can offer a novel and physically meaningful representation, setting another distinct path to further our understanding of jets and QCD in general.

3.2 Physics-Inspired Framework of Linearized Optimal Transport with Simple Machine Learning Models

The question of defining a distance between collider events is notoriously difficult to answer. Identical events at parton level can appear to differ upon reconstruction due to soft or collinear emission, while topologically distinct events at parton level can appear identical upon reconstruction, depending on the degree of coarse graining. On the other hand, as a perturbatively well-defined quantity [91], the energy flow of a jet with massless particles can be viewed as a measure, that is, a distribution of some "mass". In this case, "mass" can be either the energy or the transverse momentum of the particle. In Chapter 2, we have seen that optimal transport defines a mathematically rigorous distance between measures. It is therefore natural to test if optimal transport is up to this trying task of equipping the space of collider events with a suitable metric.

One simple OT distance—the EMD introduced in Chapter 1—has already proven valuable in numerous applications to collider physics; please refer back to Section 1.3 for a brief review. Of course, we have learnt from Chapter 2 that the EMD is but one example among a whole family of OT distances. Here we demonstrate how to apply optimal transport (including but not limited to the EMD) to collider data—more specifically, to the energy flow of an event or a jet.

In particular, we focus on the two special OT distances, W_2 and HK, and explain in full details how the LOT framework is to be coupled with simple machine learning models for the downstream classification task. However, we would also like to emphasize that the OT distances and correspondingly the LOT embeddings themselves are independent of the downstream ML models, as they simply provide the inputs. Indeed, later in Chapter 4, we will see another novel application of OT distances coupled with other statistical frameworks under an entirely different context.

3.2.1 Optimal Transport on Collider Event Energy Flow

The essential prerequisite for optimal transport computation is a ground space equipped with a ground metric, which is then lifted to a new distance defined on the space of measures. If the input ground space is not chosen wisely, then there is no reason to expect OT to output any meaningful distance between measures. Since a jet is given in the terms of its energy flow, it is natural as a first choice to pick the ground space to be the $y - \phi$ plane on the collider cylinder with an Euclidean distance same as in the jet definitions, i.e., $d_{\text{ground}} = \sqrt{\Delta y^2 + \Delta\phi^2}$.

Given that a jet consists of a finite number of constituent particles, the distributions are always sums of Dirac masses, where the "mass" of an individual particle is weighted according to its transverse momentum p_T. Furthermore, we have seen in Chapter 2 that only normalized distributions are needed in the computation of both the W_2 and the HK distances. So we always normalize all jets to have $p_T = 1$ before the OT computation.

Figure 3.1 displays the energy flows of two QCD jets as colored (one red; one blue) discrete distributions in the chosen $y - \phi$ ground space, where each point represents one constituent particle with its size proportional to its p_T (normalized by the jet's total p_T). Here we demonstrate an optimal transport plan of the W_2 distance between two jets. The darkness of the lines connecting the respective points in the two jets indicates how much p_T is moved from one particle to another.

It is important to note here that one can also consider other ground spaces. However, given the nice physical interpretation of the energy flow, we expect that a ground space built on top of it would generate a meaningful OT metric that can effectively capture the differences among a variety of jet types, as will be shown momentarily. A more general framework incorporating additional measurable quantities other than the kinematic information, such as charges and flavors, are under our current research and will be briefly mentioned in the last section of this chapter.

3.2.2 The LOT Framework for Jet Tagging

As explained in Chapter 2, for both W_2 and HK distances, one can assign a LOT coordinate to each data point once a reference point is chosen. The choice of a reference measure is not critical in practice, as long as it covers the underlying ground space in a reasonably uniform fashion. On the other hand, it is an interesting open question to consider multiple reference measures anchored on the OT manifold and see whether it brings additional information.

In our present study, we only include one reference measure, which, for the most of the times, is a discrete uniform distribution on the $y - \phi$ plane. We then use the `Python Optimal Transport` library [92] and our own custom codes to compute the OT distances

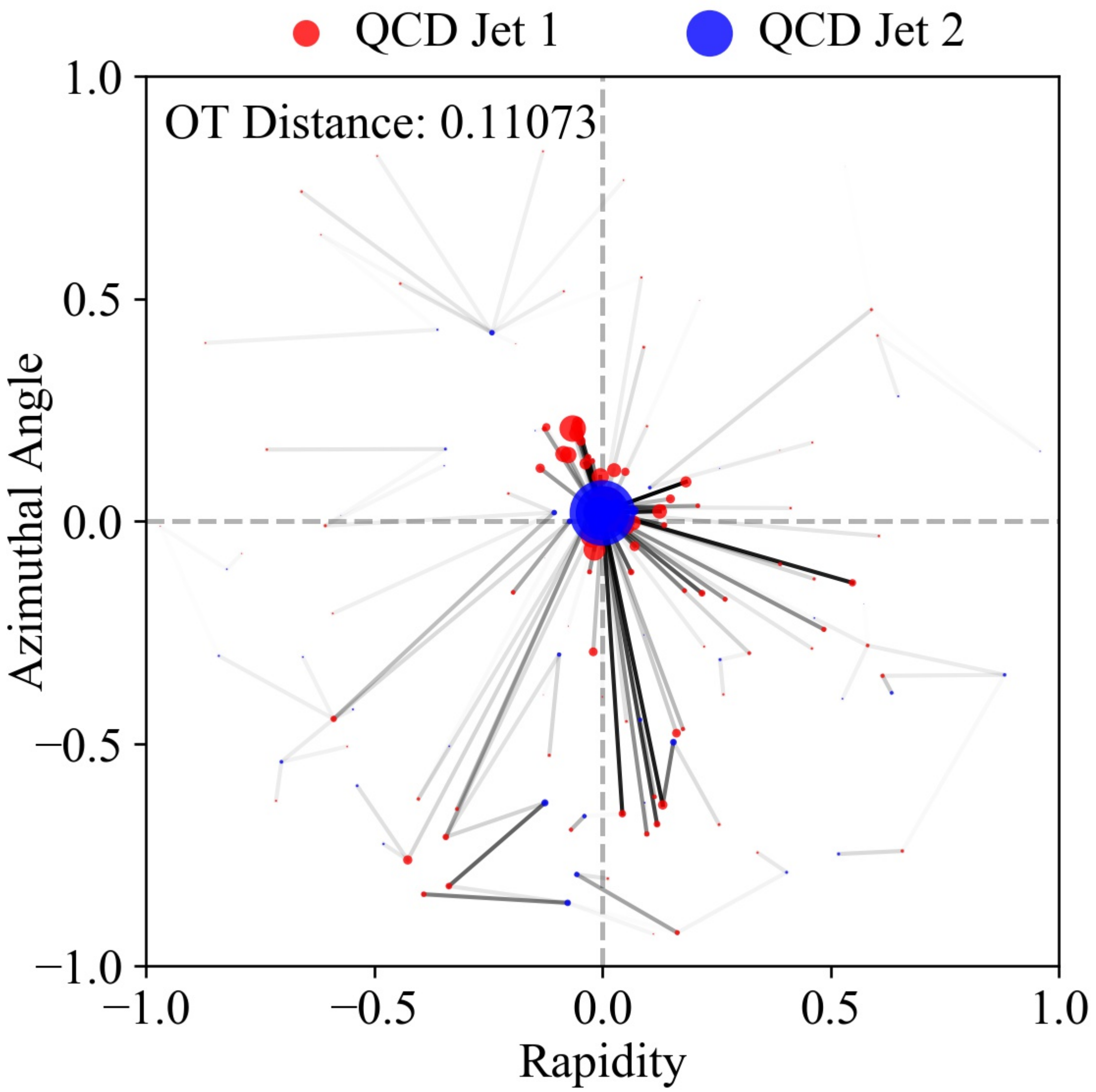


Figure 3.1: Optimal transport on the energy flows of two QCD jets (one red and the other blue) in the ground space of the $y - \phi$ plane. The OT distance used is the 2-Wasserstein distance, and the black lines show the transport plan.

between each jet in the dataset and the reference which is similarly normalized to have unit total p_T. Once we have this OT distance in hand, we proceed to calculate the linear embedding for each jet using the methods in Chapter 2. We then recover the approximate LOT pseudo-distance between any two jets from the weighted ℓ^2 distance between their LOT coordinates. In the case when our reference has all its particles with equal mass, the weighted ℓ^2 norm reduces to a classical ℓ^2 norm. This enables us to store only the jet LOT coordinates and move the actual distance computation to the downstream ML models, since now it is just the standard Euclidean distance which is implemented natively in most algorithms. To re-emphasize, the LOT coordinate is our novel representation of a collider event and constitutes the main innovation of our approach.

Now depending on our applications, we can couple a large pool of simple ML algorithms to the LOT coordinates. As our current task is jet tagging, we will mainly focus on supervised classification methods, among which k-nearest neighbor (kNN) and support vector machine (SVM) are two simple examples. But even before proceeding to classification, one would like to first take a look at the data—more specifically the LOT embedding—to gain some intuition about the OT manifold. Here, various dimensionality reduction techniques such as Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA) can aid data visualization. Further, we can also try unsupervised clustering, where we leave the model itself to assign a label for each jet. We consider the simple example of k-medoids clustering.

Although relatively limited in performance, all the aforementioned traditional ML models have important advantages over large neural networks. They are more computationally economic, have fewer hyper-parameters to tune, and offer better human interpretability. Most of them are also off-the-shelf functions implemented in the python package `scikit-learn` [93], making their adoption easier in practice. We now include a brief general description for each model. A more detailed discussion will be given when the respective performances of the models are analyzed on the actual jet tagging tasks.

Dimensionality Reduction and Data Visualization

Already in Section 2.6, we have used Principal Component Analysis (PCA) to perform a first quick look at the LOT embedding of the data. Essentially, PCA is a simple dimensionality reduction tool that finds the most important features (aka *principal components*) which maximally contribute to the overall variability of the data. This is achieved by calculating the covariance matrix of the standardized input data and finding the first few eigenvectors that correspond to the largest eigenvalues of the covariance matrix. Therefore, the first principal component, i.e., the eigenvector of the biggest eigenvalue, captures the largest amount of variability in the data, the second one captures the second largest, and so on. If the first few principal components already account for a large percentage of data variation (as is the case for the ellipses in Section 2.6), we say the data can be reduced to a much lower dimension and use the principal components to summarize and visualize the original dataset.

Closely related to PCA is another powerful and equally simple-to-implement method, Linear Discriminate Analysis (LDA) [94], which also enjoys closed-form solutions with no hyper-parameter. With the assumptions that the input data is Gaussian and the Gaussian for each class shares the same covariance matrix, LDA projects the input high-dimensional data onto a direction that is most discriminative, denoted as the LDA direction, by minimizing the data variability within each class while at the same time maximizing the separation between different classes. Of course, just like PCA, more LDA directions (i.e., discriminants) can be retained. Since here we are exclusively concerned about the binary classification problem, we only need one LDA direction to separate data into the signal class and the background class. Later we will use LDA both as a tool for visualization and as a classifier.

Supervised Classification

The classifier k-nearest neighbor (kNN) [95] relies on a simple majority vote of one's closest k neighbors in the training set to determine the class membership of the new data point. Here k is a model hyper-parameter to be tuned. Since kNN relies only on a notion of pairwise distances, it serves as a good probe to check whether our LOT approximation sufficiently captures the difference among various jet types while at the same time adequately reflecting the similarity within one specific type. The simplicity in understanding kNN and its reliance only on pairwise distances between events contribute to its adoption in the original EMD paper [4] and our own studies.

The support vector machine (SVM) [96] is slightly more sophisticated. Essentially, SVM lifts the inputs into a high-dimensional space and finds an optimal hyperplane to best separate the data. Key to SVM is the choice of a kernel function. We use the common `rbf` kernel, i.e., $\exp[-\gamma d(x,x')^2]$ where $d(x,x')$ is the LOT pseudo-distance between the two data points. Here γ is a tunable hyper-parameter controlling how much influence a single training example has. A high γ suggests that only nearby points are considered. The other hyper-parameter C regulates the strength of the penalty term when a sample is misclassified, with a high C implying that nearly all training examples need to be classified correctly.

Unsupervised Clustering

For unsupervised learning, we choose as a first try k-medoids clustering [97] implemented in the python package `pyclustering` [98]. The goal of k-medoids clustering is to partition the dataset so that the distance between points labeled to be in a cluster and the point designated as the center of that cluster is minimized. Note that the centers, called medoids, are chosen from actual data points. For our application, the model is asked to group the unlabeled data into $k = 2$ clusters. Afterwards, the true labels are uncovered. The cluster with a higher percentage of signal jets is denoted as the signal cluster, whereas the other is designated as the background cluster.

We also retrieve the true labels of the two picked medoids. Ideally, the true label of the medoid should be the same as the label of its own cluster. If not, we prefer the cluster's label. We then assign all jets in the signal cluster as signals, and those in the background cluster as background jets. This assignment is compared with the ground truth to assess the performance of our clustering model. Strictly speaking, the model is semi-supervised, for we need the true labels to decide which cluster is the signal cluster.

3.3 Jet Tagging with Balanced Optimal Transport

Here we focus on the task of jet tagging using the balanced 2-Wasserstein distance and its linearization scheme. The results presented are based on [35]. We consider in total five types of jets: single-pronged QCD (quark and gluon) jets, two-pronged boosted W boson jets, three-pronged boosted top quark jets, two-pronged boosted Higgs boson jets, and two-pronged boosted jets from a hypothetical new particle ϕ. This new BSM particle ϕ is taken to be a scalar transforming in the $\mathbf{6}$ representation of $SU(3)_C$ and carrying electromagnetic charge $+\frac{1}{3}$; we consider a benchmark mass of $m_\phi = 100$ GeV with a width of $\Gamma_\phi = 2$ GeV. It couples equally to all quark pairs that respect charge conservation. We calculate the Feynman rules for this BSM particle ϕ using FeynRules [99].

Instead of examining all possible pairwise combinations, we narrow our analysis to the following seven pairs: W *vs* QCD, t *vs* QCD, t *vs* W, H *vs* QCD, H *vs* W, BSM *vs* QCD, and BSM *vs* W. For the most part, these comparisons could be thought of as treating both QCD and W boson jets as backgrounds, whereas top, Higgs boson, and

BSM jets are treated as signals. The W *vs* QCD pair is introduced as a benchmark for the performance of the other six tagging tasks, as well as for a meaningful comparison with the results obtained in [4].

We generate proton-proton collision events using MADGRAPH 2.6.7 [99] at $\sqrt{s} = 14$ TeV, where the two-pronged boosted Higgs boson jets are generated via $q\bar{q} \to Z(\to \nu\bar{\nu}) + H(\to b\bar{b})$, and the BSM jets through $q\bar{q} \to \phi\bar{\phi}$; all other SM jets are created via pair production. The BSM (anti)particle subsequently decays to two quarks. The matrix elements are then fed into PYTHIA 8.243 [100], with hadronization and multiple particle interactions switched on using default tuning and showering parameters. No detector simulation is included. Afterwards, we cluster the jets in FASTJET 3.3.2 [84] using the anti-k_T algorithm with a jet radius of $R = 1.0$, where at most two jets with $p_T \in [500, 550]$ GeV and $|y| < 1.7$ are kept.

To remove any artificial difference in the energy flows of the produced jets, every jet is preprocessed by boosting and rotating to center the jet four-momentum and vertically align the principal component of the constituent p_T flow in the rapidity-azimuth plane using the `EnergyFlow` package [90, 87, 101, 4, 14]. This preprocessing step is essential to make sure that the difference between two jets' energy flows depends on their internal substructural distinction, not on some overall transformation irrelevant to the underlying physics. However, it should be noted that the jet preprocessing scheme is not uniquely defined or physically well-grounded. Further, there exists no general consensus on how to preprocess an entire collider event. The hope is that an upgraded version of the OT framework can circumvent such ambiguities arising from different preprocessing procedures; see Section 3.7.

3.3.1 Linearized W_2 Embedding for Jets

Though not necessary, we work with a single choice of the reference measure in order to have a unified framework for the seven comparison tasks. Our reference jet has a total p_T of 525 GeV and 225 constituent particles, each with the same amount of p_T evenly distributed on a 15×15 grid with $|y| \leq 1.7$ and $|\phi| \leq \frac{\pi}{2}$. This corresponds to an isotropic distribution on the cylinder; note that related reference distributions were explored in [15] for the purposes of defining the event isotropy variable. We have also tried other reference jets and the resulting LOT approximation does not show any material difference compared to what is obtained from the uniform reference jet. Furthermore, as we justify rigorously in Chapter 2, the $\mathrm{LinW}_{2,\mathcal{R}}$ approximation with a uniform reference measure $\mathcal{R}$ can be seen as an approximation of the actual LinW_2 metric, which approximates the original 2-Wasserstein metric at large and small distances; see Equation (2.53).

Figure 3.2 shows the optimal energy movements between two sample QCD jets and between QCD and W jets using the exact W_2 distance and its linear approximation, respectively. In visualizing the $\mathrm{LinW}_{2,\mathcal{R}}$ pseudo-distance, vectors located at each particle in the reference jet indicate the *difference* between movement of p_T from that particle in the reference jet to particles in the respective sample jets. In each case, the total distance between the two jets is also shown.

These examples illustrate the qualitative properties of both distances applied to simulated events: in the case of W_2, large OT distances correspond to the movement of significant amounts of energy between particles widely separated in the ground metric, while large $\mathrm{LinW}_{2,\mathcal{R}}$ pseudo-distances correspond to very different transport plans between the reference jet and the respective particles. We observe that the $\mathrm{LinW}_{2,\mathcal{R}}$ pseudo-distance is numerically close to the exact W_2 distance, consistent with our expectation.

In principle, the $\mathrm{LinW}_{2,\mathcal{R}}$ pseudo-distance, the LinW_2 metric, and the W_2 metric do

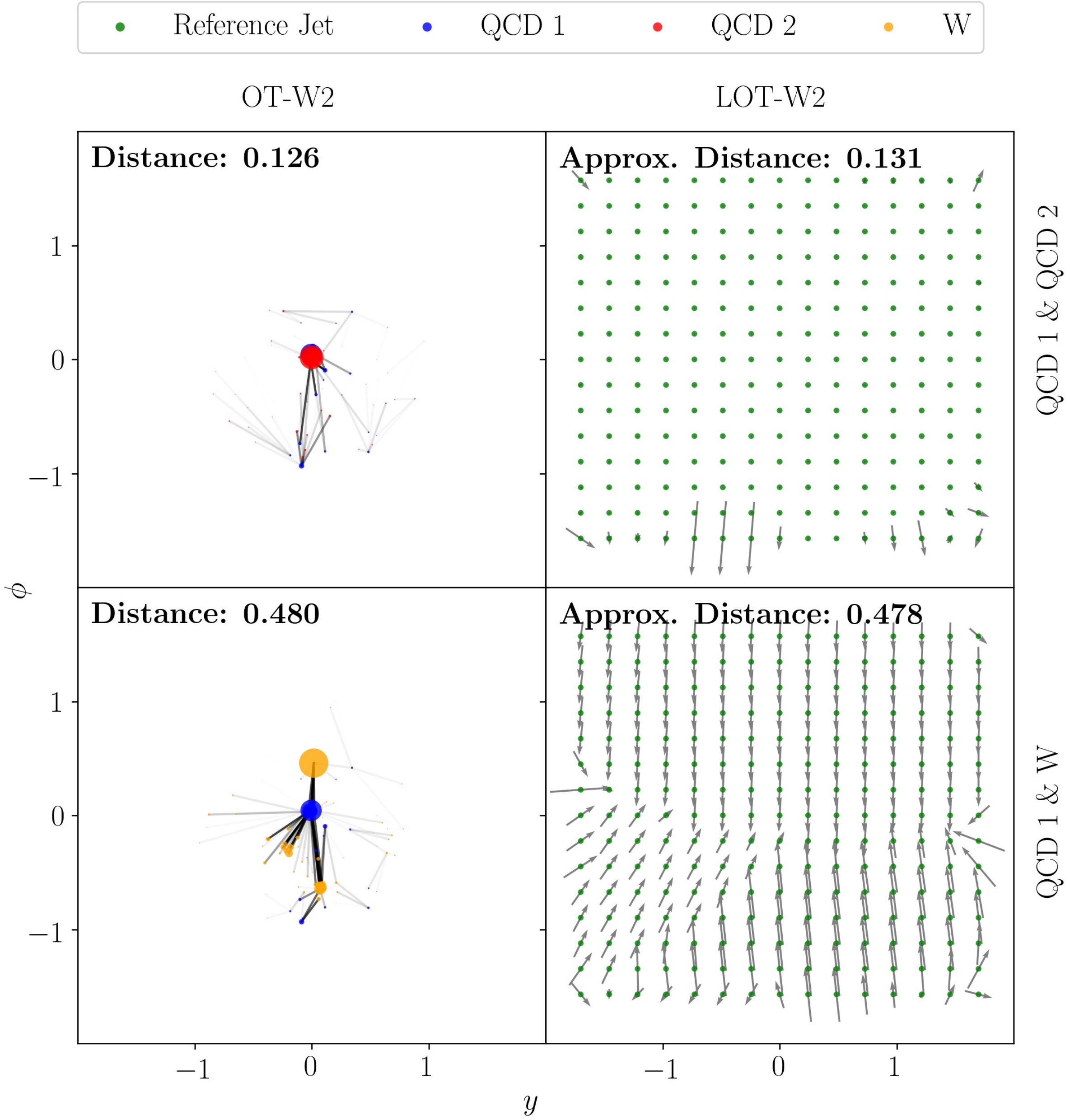


Figure 3.2: *Upper left:* The optimal movement to rearrange one QCD jet (red) into another (blue) using the exact W_2 metric (denoted as OT-W2). *Upper right:* The optimal movement to rearrange the same two QCD jets using $\mathrm{LinW}_{2,\mathcal{R}}$ (denoted as LOT-W2). *Lower left:* The optimal movement to rearrange a W jet (orange) into a QCD jet (blue) using the exact W_2 metric. *Lower right:* The optimal movement to rearrange the same QCD and W jets using $\mathrm{LinW}_{2,\mathcal{R}}$.

not need to attain similar values in order for LOT to offer good discrimination power in classification and clustering tasks. However, as an illustration of the similarity of $\mathrm{LinW}_{2,\mathcal{R}}$ and the W_2 metric in practice, we plot in Figure 3.3 a histogram of the difference between the $\mathrm{LinW}_{2,\mathcal{R}}$ approximation and the exact W_2 metric when computing the pairwise distances between a sample of 500 mixed W and QCD jets. We observe that the $\mathrm{LinW}_{2,\mathcal{R}}$ is on average slightly larger than W_2 (mean 0.67%), and they are generally of comparable size (standard deviation = 5.82%).

3.3.2 Tagging Results

For every comparison task, we create two balanced datasets, each with about 50% signal jets. The smaller one, named the sample dataset, consists a total of 10,000 jets and is mainly used for picking the best hyper-parameters, though it also constitutes a complete analysis in its own right. The full dataset has 140,000 jets in total, and is used to assess the model performance and draw the final conclusions.

For the two classifiers kNN and SVM, the sample dataset is further divided into a training sample of 5000 jets, a validation sample of 2500 jets used to decide the best hyper-parameters, and a test sample of 2500 jets. The full dataset is split into a training set of 100k jets and a test set of 40k jets for these two models. For kNN, we test the hyper-parameter k in the range from 10 to 1000 with an increment of 10, whereas for SVM the hyper-parameters C and γ both run from 10^{-5} to 10^{5} again with an increment of 10. Thus, SVM needs to be run for $11 \times 11 = 121$ times to determine the best choice of the (C, γ) pair.

For LDA, thanks to its high efficiency, we train and test on both the sample dataset (training sample size = 8000, test sample size = 2000; validation sample is not needed since there's no hyper-parameter for LDA) and the full dataset (training set size = 100k,

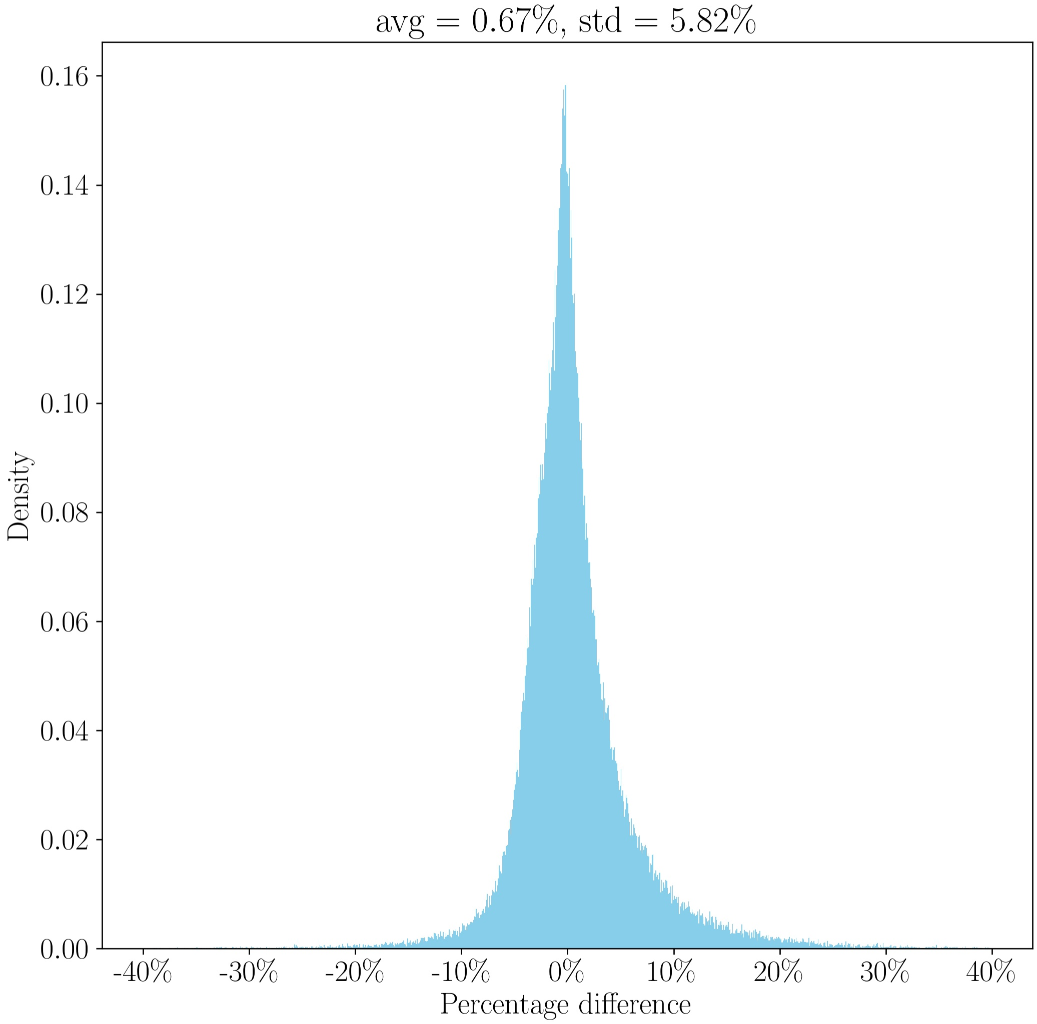


Figure 3.3: Distribution of percentage differences between the $\mathrm{LinW}_{2,\mathcal{R}}$ pseudo-distance and the W_2 distance for pairs of events in a sample of 500 mixed W and QCD jets.

test set size = 40k), which amounts to two separate, identical analyses with different sizes. The k-medoids algorithm has only been applied to the sample dataset due to its computational intensity, and in this case, all 10,000 jets are fed into the model at once for clustering.

Figure 3.4 displays the receiver operating characteristic (ROC) curves of the three classifiers kNN, SVM and LDA for each of the seven comparison tasks. Also included is the Area Under the ROC Curve (AUC) which encapsulates the model performance in a single number between 0 and 1. An AUC close to 1 is most desirable, whereas a value around 0.5 suggests a random classifier, the worst-case scenario. All results are obtained on the full test datasets consisting of 40k jets, using the models trained on 100k jets with hyper-parameters, if present, picked by the sample datasets.

To get a better sense of the model performance, we compare the AUCs of our LOT-coupled ML models for the W *vs.* QCD classification task with other common classifiers built in [4] where the training set, though different, also contains 100k balanced W and QCD jets, and the test set contains 20k such jets. The model most akin to our $\mathrm{k}_{=20}$NN-LOT is $\mathrm{k}_{=32}$NN-EMD built upon the EMD proposed in [4], which is an interpolation between the exact W_1 distance and total variation norm. [1] The other benchmarks include the N-subjettiness ratio $\tau_2^{\beta=1}$ / $\tau_1^{\beta=1}$, the Energy Flow Network (EFN) and Particle Flow Network (PFN) neural networks [87], and a linear classier trained on Energy Flow Polynomials (EFPs) [90]; please refer to the original papers for more details.

[1] Although our samples are not identical to those in [4], we apply the same prescription for simulating and preparing the samples, and our W/QCD jet samples yield results for $\mathrm{k}_{=32}$NN-EMD compatible with [4].

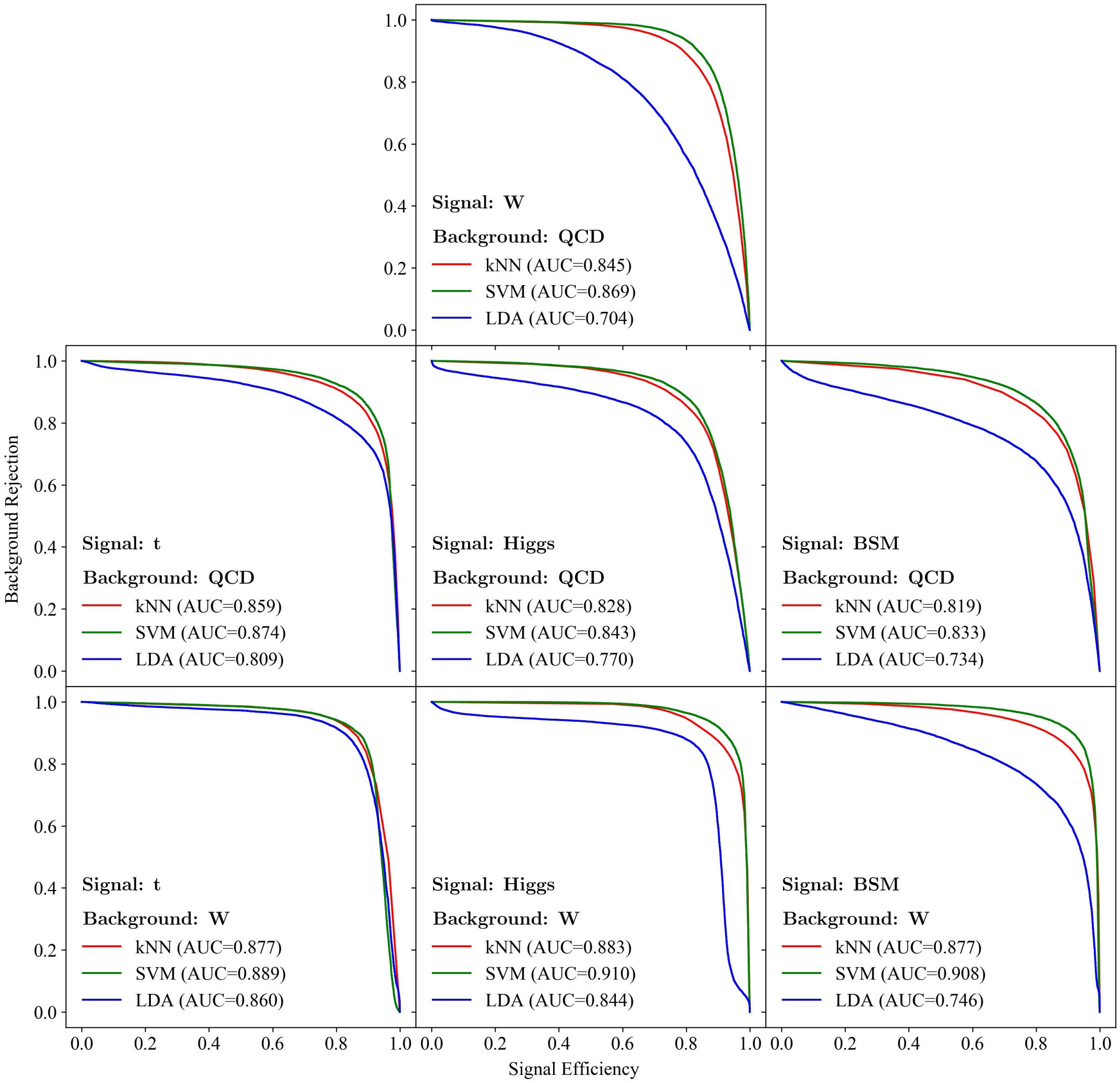


Figure 3.4: ROC curves for the seven jet tagging tasks evaluated on the full test datasets of 40k jets. The x coordinate shows the signal efficiency rate and the y coordinate gives the background rejection rate.

Datasets	**Model**	**AUC**
Our Datasets	k$_{=20}$NN-LinW$_{2,\mathcal{R}}$	0.845
	SVM-LinW$_{2,\mathcal{R}}$	0.869
	LDA-LinW$_{2,\mathcal{R}}$	0.704
Datasets in [4]	k$_{=32}$NN-EMD	0.887
	$\tau_2^{\beta=1}$ / $\tau_1^{\beta=1}$	0.776
	PFN	0.919
	EFPs	0.917
	EFN	0.904

Not surprisingly, the neural networks obtain the best performance. But the four optimal transport inspired models (three with LinW$_{2,\mathcal{R}}$ and one with EMD) are on a par with these state-of-the-art complex classifiers, and they significantly outperform the N-subjettiness observable (with the single exception of the exceptionally simplistic LDA).

More pertinent to our current investigation is the observation that models coupled with the LinW$_{2,\mathcal{R}}$ approximation perform as well as those using the exact EMD metric. The AUCs of kNN-LinW$_{2,\mathcal{R}}$ and SVM-LinW$_{2,\mathcal{R}}$ are close to the AUC of kNN-EMD, suggesting that it does not make much difference for jet tagging whether we use the exact OT metric or its linearized version.

Yet on the practical level, the LinW$_{2,\mathcal{R}}$ approximation has a significant advantage over the exact EMD metric. The computation of the LinW$_{2,\mathcal{R}}$ coordinates for 140k jets only takes about 10 minutes on a desktop computer, whereas it is infeasible to compute the full exact EMD matrix of pairwise distances on the same computer and still requires significant time on a cluster.

Table 3.1 summarizes the results obtained for all seven comparison tasks, with complete, independent analyses done both on the sample datasets and the full datasets. In addition to AUC, we also report the True Positive Rate (TPR) and False Positive Rate (FPR), where the TPR is the same as the signal efficiency, and the FPR equals to one minus the background rejection. A TPR near 1 and a FPR close to 0 are preferable. For SVM and kNN, we also include the hyper-parameters chosen by the sample datasets.

The results for k-medoids are harder to interpret, so we defer a full discussion to a later section.

Also included in the table is the approximate run time for each task, performed on an iMac with 3.6 GHz 8-Core Intel Core i9 and 16 GB memory. The longest analysis takes no more than 10 hours, which, when combined with the extra few minutes for calculating the $\mathrm{LinW}_{2,\mathcal{R}}$ coordinates, is quite manageable. LDA in particular only takes seconds to process the full datasets and in this light its classification results are surprisingly good. In addition, models performed on the sample datasets require as few as 2 hours for a full scan of hundreds of possible combinations of hyper-parameters. Competitive classification performance coupled with efficient computational time suggests that our LOT framework plays an effective role in jet tagging alongside the exact OT metric, complex neural networks, and traditional handpicked observables.

Given that the sample datasets constitute complete analyses in their own rights, we can compare their results with those obtained using the full datasets. In general, model performance naturally gets better with more training data, but we observe that the increase in performance going from 10k jets to 140k jets is perhaps not significant enough to justify the extra computational resources needed. Since the numbers quoted for AUC, TPR and FPR are only intended as general performance evaluations rather than precise measures, the fluctuations in these numbers can be safely ignored and we therefore conclude that a dataset of 10,000 jets (with as few as five thousands for training) is already enough to assess the overall quality of the model and the underlying metric.

Table 3.1: Results for the seven jet tagging tasks using four different machine learning models coupled with the $\mathrm{LinW}_{2,\mathcal{R}}$ embedding.

Model	Dataset		Comparison Task						
			W *vs* QCD	t *vs* QCD	t *vs* W	H *vs* QCD	H *vs* W	BSM *vs* QCD	BSM *vs* W
LDA	Sample Dataset	**AUC**	**0.6896**	**0.7863**	**0.8464**	**0.7642**	**0.7865**	**0.7158**	**0.7244**
		TPR	0.6926	0.7746	0.7886	0.7378	0.7762	0.6713	0.6562
		FPR	0.3133	0.2020	0.0958	0.2095	0.2032	0.2397	0.2074
		Approx. Run Time	several seconds						
	Full Dataset	**AUC**	**0.7041**	**0.8077**	**0.8573**	**0.7703**	**0.8443**	**0.7337**	**0.7455**
		TPR	0.7156	0.7969	0.7957	0.7661	0.8254	0.7549	0.6804
		FPR	0.3075	0.1815	0.0812	0.2255	0.1368	0.2874	0.1894
		Approx. Run Time	several seconds						
SVM	Sample Dataset	**AUC**	**0.8410**	**0.8630**	**0.8751**	**0.8349**	**0.8831**	**0.8239**	**0.8806**
		TPR	0.8148	0.8929	0.8333	0.8006	0.8750	0.8582	0.9090
		FPR	0.1327	0.1669	0.0831	0.1308	0.1088	0.2104	0.1478
		Approx. Run Time	2 hours						
	Full Dataset	**AUC**	**0.8687**	**0.8780**	**0.8805**	**0.8426**	**0.9100**	**0.8331**	**0.9077**
		TPR	0.8451	0.8873	0.8365	0.8185	0.9103	0.8471	0.9191
		FPR	0.1077	0.1313	0.0755	0.1332	0.0904	0.1808	0.1037
		Approx. Run Time	6 hours						
	Hyperparameters	C	1.0	1.0	10.0	1.0	1.0	1.0	1.0
		γ	100.0	100.0	10.0	100.0	100.0	100.0	100.0
kNN	Sample Dataset	**AUC**	**0.8191**	**0.8450**	**0.8659**	**0.8203**	**0.8628**	**0.8026**	**0.8361**
		TPR	0.7741	0.8164	0.8040	0.7975	0.8295	0.8172	0.8241
		FPR	0.1358	0.1264	0.0723	0.1568	0.1038	0.2120	0.1520
		Approx. Run Time	15 minutes						
	Full Dataset	**AUC**	**0.8455**	**0.8601**	**0.8735**	**0.8280**	**0.8831**	**0.8192**	**0.8772**
		TPR	0.8033	0.8217	0.8156	0.8040	0.8566	0.8261	0.8836
		FPR	0.1123	0.1014	0.0686	0.1479	0.0905	0.1876	0.1292
		Approx. Run Time	4 hours						
	Hyperparameter	k	20	40	10	20	20	10	20
k-medoids Clustering	Sample Dataset	**AUC**	**0.6797**	**0.8096**	**0.8074**	**0.7689**	**0.8028**	**0.7622**	**0.6698**
		TPR	0.7947	0.9282	0.6583	0.8374	0.6835	0.8837	0.5216
		FPR	0.4354	0.3089	0.0436	0.2996	0.0778	0.3592	0.1821
		Signal Percentage (sig, bkg)	(63.78%, 25.97%)	(74.70%, 9.27%)	(94.00%, 27.02%)	(73.60%, 18.81%)	(90.11%, 26.24%)	(71.05%, 15.33%)	(74.81%, 37.75%)
		Clusters' Size (sig, bkg)	(6118, 3882)	(6159, 3841)	(3565, 6435)	(5682, 4318)	(3861, 6139)	(6211, 3789)	(3549, 6451)
		Medoids True Labels (sig: 1, bkg: 0)	(1, 0)	(0, 0)	(1, 0)	(1, 0)	(1, 1)	(1, 0)	(1, 0)
		Approx. Run Time	30 minutes						

3.3.3 Discussions

Some general features can be immediately read off from Table 3.1. Whichever jets we compare, SVM always gives the best classification performance with AUCs around 0.9, approaching the performance of neural networks. This suggests that jets represented in their $\mathrm{LinW}_{2,\mathcal{R}}$ coordinates are indeed very well separated by a hyperplane in some high-dimensional feature space, which in turn demonstrates the fitness of the approximate metric itself. Except for t *vs.* W jets classification, the hyper-parameters chosen for SVM via the validation process are all the same, with $C = 1$ and $\gamma = 100$. It means that the model uses only a reasonable amount of regularization and thus a relatively smooth decision surface is drawn. On the other hand, a γ of 100 is considered large, indicating that only nearby samples can have an influence on the classification of a new point.

This latter observation is consistent with what is suggested by the hyper-parameter k picked by kNN. All seven comparison tasks prefer small k values less than 50, which means that to determine the type of an unknown jet we need to look no further than its closest 50 neighbors. If W_2 does not place same-type jets near each other as desired, then models with hyper-parameters preferring locality won't be able to achieve such satisfying classification performances. Therefore, the hyper-parameters picked by SVM and kNN provide an indirect evidence for the suitability of the optimal transport metric—it indeed groups jets of the same type near each other and separates those of different types. We will later turn this speculation into more convincing and intuitive visualization.

Among the seven jet tagging tasks, kNN and SVM both have the best performance in distinguishing Higgs boson jets from W boson jets and are least capable of separating BSM jets from QCD jets. This is mainly caused by a relatively high false positive rate, meaning that the models have a tendency to wrongly classify QCD jets as BSM jets. The same reason applies to LDA when it performs poorly on W *vs.* QCD classification relative to other tasks. For each type of signal jets (t, H, or BSM), all three classification models perform better when the background is W jet rather than QCD jet.

Unsupervised Clustering for Jets

We now focus on the k-medoids clustering algorithm, which is only analyzed on the sample datasets due to computational limitations. Given that unsupervised learning is inherently more difficult than supervised learning, it's not surprising to see the performance of k-medoids algorithm to be inferior to that of kNN or SVM. But even then, except for the W *vs.* QCD and BSM *vs.* W tasks, the AUCs of k-medoids are all above 0.75, on a par with the supervised learning models analyzed on the sample datasets. The clustering algorithm even shows superior performance compared to LDA for most tagging tasks. This remarkable achievement again points to the merit of the underlying 2-Wasserstein distance and is encouraging for the further exploration of optimal transport applications to unsupervised learning algorithms.

It should be noted that AUC is not the only gauge of model performance. Especially in the case of k-medoids clustering, we also need to take a look at other indicators to map a more complete picture. Beside examining the TPR and FPR, we also like to know more about the properties of the two clusters outputted by the algorithm. If the model is perfect, then each cluster should contain only signal jets or only background jets. The purity of the two clusters is given in the second row of k-medoids clustering in Table 3.1, where we record the signal percentage (defined as the number of signals in the cluster divided by the total number of jets in that cluster) in the signal cluster and the background cluster, respectively.

By definition, the signal cluster is the group with a majority of signal jets, which, if pure, should have a signal percentage of 100%. Similarly, a pure background cluster should have 0% signal percentage. Notice that the sum of the signal percentage of the two clusters does not necessarily equal to 1 (but in the ideal case it is). The worst-case scenario is to have the signal percentage of both clusters close to 50%. A quick look at the second row at least qualitatively confirms that the AUC of the task is indeed higher whenever we have two purer clusters, with the best AUC obtained for t *vs.* QCD clustering which has a signal percentage of 74.70% for the signal cluster and only 9.27% for the background cluster.

The size of the clusters also reveals how well the model performs. Ideally, the result would be two clusters with equal size, that is, each with about 5000 jets, since the data itself is balanced. Here the best result we have is for H *vs.* QCD task, where the Higgs cluster has 5682 jets and the QCD cluster has a total of 4318 jets. But in general, the two clusters are not well balanced. In the worst case, the W cluster has 81.77% more jets than the BSM cluster, and it does correspond to the lowest AUC score.

In theory, the two medoids should be the most representative jet for the clusters they respectively belong. Since the medoids are actual data points, we can uncover their true labels and check whether they agree with the type of the cluster they are assigned to. Only the two tasks, t *vs.* QCD and H *vs.* W, give conflicting answers. For the t *vs.* QCD clustering, the two chosen medoids are both background QCD jets. Thus the signal top cluster acquires a QCD jet as its representative. The situation is reversed for the H *vs.* W task where now the background W cluster elects a signal Higgs jet as its exemplar. Nevertheless, both tasks enjoy high AUC scores, which suggests that the true labels of the medoids might not have a direct influence on model performance.

The general message here is that AUC, though powerful and straightforward, is not enough to assess the performance of an algorithm. Other indicators are required to gain a fuller appreciation of the strength and weakness of the model, both for clustering and for classification.

Visualization of the LOT Manifold

We use LDA to visualize and aid understanding of the LOT approximation of the W_2 metric and its associated Euclidean embedding. Given the 225×2 linearized coordinate for each jet, we first stack the list of the second coordinate ϕ at the end of the list of the first coordinate y and reshape the coordinate to be 450×1, which is then fed into a LDA model for the projection of the 450 coordinates onto one single most discriminative direction (denoted as the LDA direction). This allows us to represent every jet as one single point on the LDA direction for easy visualization.

Figure 3.5 shows such projection for the 10000 jets in the t *vs.* W sample dataset, which enjoys the highest AUC among the seven tasks with the LDA classifier. A clear separation between W and top jets can be seen, with the majority of W boson jets grouped towards the left end of the LDA direction and most top jets towards the right end, explaining the good performance of the LDA classifier for this task.

It is enlightening to see how jets vary along the chosen LDA direction. To this end, we first select the jet whose 1-dimensional projected LDA coordinate has a value closest to the mean of all LDA coordinates in the dataset and denote it as the mean jet. We then compute the standard deviation of the dataset. Now jets whose LDA coordinates are up to 3 sigmas away from the mean jet are displayed in Figure 3.5. We observe a clear tendency of particles spreading more on the $y - \phi$ plane as we move from the left end of the LDA direction to the right end, i.e., from negative sigmas to positive sigmas, corresponding well to our intuition that top jets are more smeared and tend to have a three-pronged structure.

As another illustration, we examine more closely how the W_2 metric rearranges the p_T of one jet to make it look like another, as shown in Figure 3.6. Here we first select the rightmost top jet t^1 and the leftmost W boson jet W^1 in the bottom plot of Figure 3.5. We

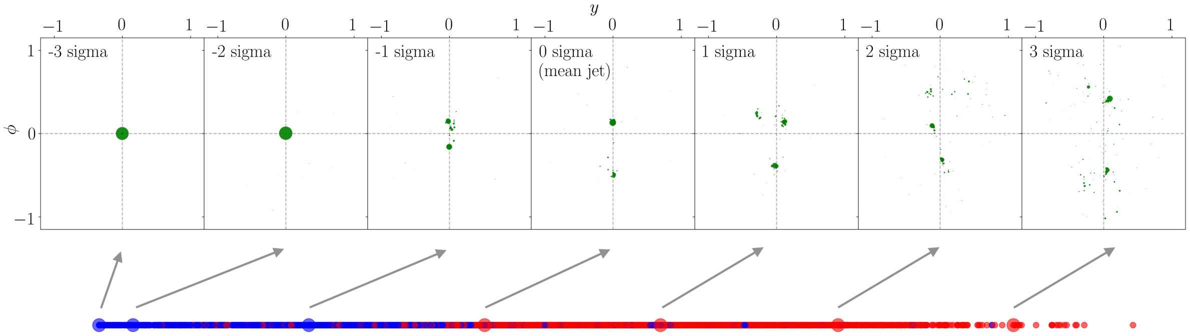


Figure 3.5: *Bottom:* Projection of the $\mathrm{LinW}_{2,\mathcal{R}}$ coordinates of 10,000 jets in the sample dataset onto the LDA direction chosen by the model. Blue dots represent W boson jets and red dots refer to top jets. The seven larger dots represent jets whose LDA coordinates are $-3, -2, -1, 0, 1, 2, 3$ sigma away from the mean jet (starting from the left). *Top:* The energy flow in the rapidity-azimuthal plane of the seven jets chosen in the bottom plot respectively. The intersection of the dashed lines shows the location of the origin in the y-ϕ plane.

then compute the exact 2-Wasserstein optimal transportation matrix γ_{ij}, which instructs how much of p_T is moved from particle i in jet W^1 (denoted as W_i^1) to particle j in jet t^1 (denoted as t_j^1). To interpolate between the two extreme jets, we create a new jet that depends on an interpolation parameter $\alpha \in [0,1]$, where $\alpha = 0$ outputs a jet identical to W^1 and $\alpha = 1$ recovers the t^1 jet. This new artificial jet$^\alpha$ contains $i \times j$ particles, each with

$$\begin{aligned} p_T^\alpha &= \gamma_{ij}, \\ y^\alpha &= (1-\alpha) \times y(W_i^1) + \alpha \times y(t_j^1), \\ \phi^\alpha &= (1-\alpha) \times \phi(W_i^1) + \alpha \times \phi(t_j^1), \end{aligned} \tag{3.10}$$

where $y(W_i^1)$ is the y coordinate of the ith particle in jet W^1, and likewise for the others. From the perspective of optimal transport theory, this artificial jet is precisely the 2-Wasserstein geodesic between the jets. Several values of α are picked in Figure 3.6 so as to show a few representatives of the interpolated jets and help us to understand intuitively

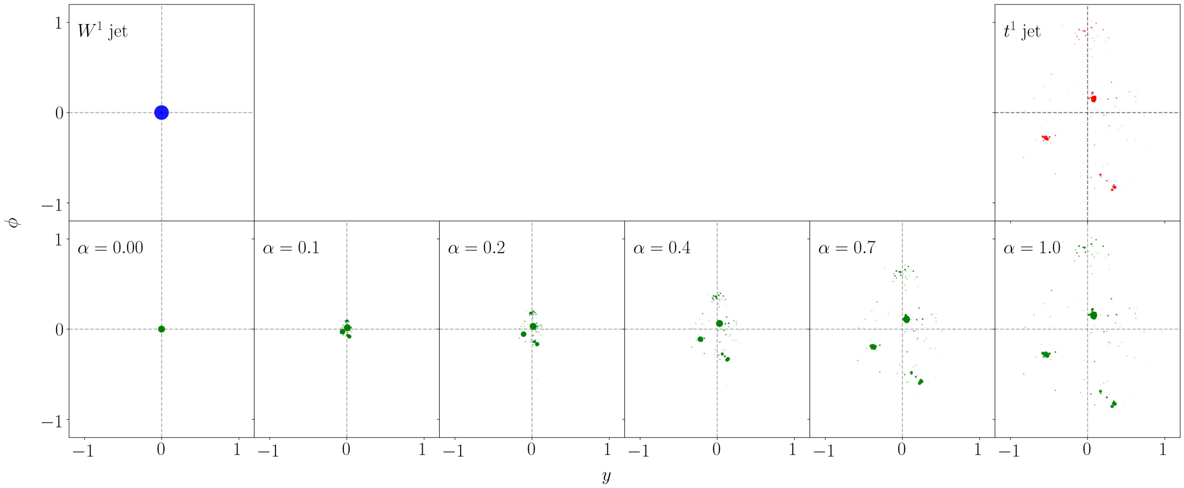


Figure 3.6: The W_2 movement of p_T to rearrange the leftmost W boson jet W^1 (blue) into the rightmost top jet t^1 (red) in the sample dataset. The intermediate green plots show artificial jets created via the interpolation parameter α. When $\alpha = 0$ and 1, the jets are respectively identical to W^1 and t^1 up to visualization. Again the intersection of the dashed lines shows the location of the origin.

the p_T movement by the W_2 metric. This interpolation technique may prove relevant to the fast simulation of collider events, insofar as it allows interpolation between real events.

3.4 Jet Tagging with Unbalanced Optimal Transport

We now extend to the unbalanced Hellinger-Kantorovich metrics, where we use the same simulated dataset as in the previous section but focus exclusively on the W *vs.* QCD tagging task with 10k jets in total. The same analysis has been extended to other pairwise tagging tasks with qualitatively similar results obtained. We therefore omit a repetitive discussion. All the results presented below are based on Section 5.4 in [34].

The analysis pipeline for $\mathrm{LinHK}_{\kappa,\mathcal{R}}$ embedding pretty much parallels that of $\mathrm{LinW}_{2,\mathcal{R}}$ embedding. The reference measure is again chosen as a regular Cartesian grid of 15 × 15 points covering the rectangle $[-1.7, 1.7] \times [-\pi/2, \pi/2]$ with uniform p_T distribution.

There is, however, one key distinction. As we have seen in Chapter 2, when using HK, the choice of the length scale parameter κ is crucial for the success of the analysis. This hyperparameter κ complicates the HK analysis and cannot be determined a priori. Therefore, we scan values over several orders of magnitude, i.e. $\kappa \in [0.01, 100] \cup \{+\infty\}$, where the W_2 distance is denoted by $\kappa = +\infty$.

Figure 3.7 displays the various optimal transport distances, including W_2 and HK, from the uniform reference measure to a QCD and a W jet, respectively. Considering the intrinsic length scale of the sample and reference measures themselves, we expect that $\kappa = 100$ is very close to balanced W_2 and $\kappa = 0.01$ essentially behaves like the Hellinger distance. In particular, for the latter, the maximal transport distance is substantially smaller than the grid spacing of the reference jet (≈ 0.2). Therefore, virtually no transport happens and almost all mass is being created and destroyed, which is why $\kappa = 0.01$ was excluded from Figure 3.7 as it looks identical to the $\kappa = 0.1$ case to naked eyes. Consequently, one would expect the classification performance to deteriorate for $\kappa = 0.01$, which will be confirmed later. Our hope is to observe boosted performance in the regime where both transport and mass creation/destruction are relevant, i.e., κ roughly between 0.1 and 1, which would then justify the introduction of this more complicated HK distances.

3.4.1 Visualization of the OT Manifolds

To gain more intuition about the various OT distances, we first apply the dimensionality reduction techniques to visualize the different OT manifolds under consideration.

We start with a principal component analysis. The jet dataset exhibits a high intrinsic dimensionality. Approximately 30 modes are needed to capture at least 90% of the dataset variance (W_2: 27; $\mathrm{HK}_{\kappa=10}$: 28; $\mathrm{HK}_{\kappa=1}$: 38; $\mathrm{HK}_{\kappa=0.1}$: 26). Moreover, for all κ

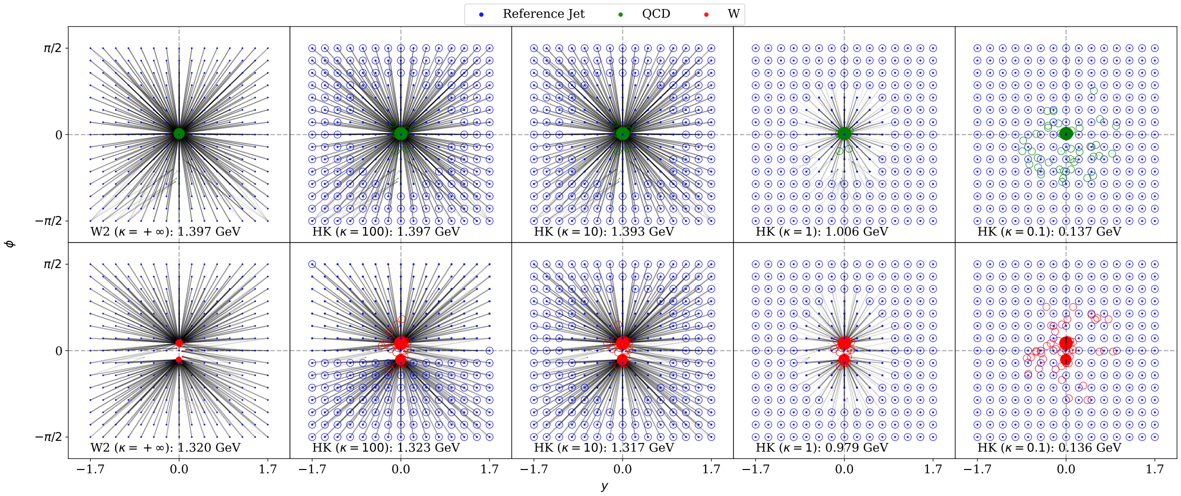


Figure 3.7: Optimal transports between the uniform reference measure (blue) and a typical QCD jet (green, first row), or a W jet (red, second row), using W_2 and HK with $\kappa = 100, 10, 1, 0.1$ (from left to right with $\kappa = +\infty$ denoting balanced W_2). The darkness of the lines indicates how much p_T is moved from one particle to another. For HK, the thickness of the circles around the points represents how much p_T is destroyed for that particular particle. Shown at the bottom of the plots are the total OT distances between the jets, which are similar for $\kappa = +\infty, 100, 10$, the transport regime. Figure copied from [34].

(including $\kappa = +\infty$), the mean of the sample point cloud is far from any sample, which indicates that the samples lie on a submanifold of the tangent space with non-trivial shape and topology. Therefore we do not find it surprising that applying the exponential map to individual dominant modes relative to the mean, or projecting samples to very few modes, do not yield artificial jet images useful for physical interpretation.

Figure 3.8 plots the the distribution of the dataset in the tangent space with respect to the first two dominant modes for the case when $\kappa = 1$. We observe that the two classes are discernible as distinct clusters with relatively little overlap and that the two coordinate values are highly dependent. In addition, for several points in the PCA coefficient space we show the actual jets in the dataset that are closest to those coefficients, as well as the approximated jet generated by the exponential map at a chosen simulated jet (jet 2 in the plot). This indicates that variations of the PCA coefficients correspond to

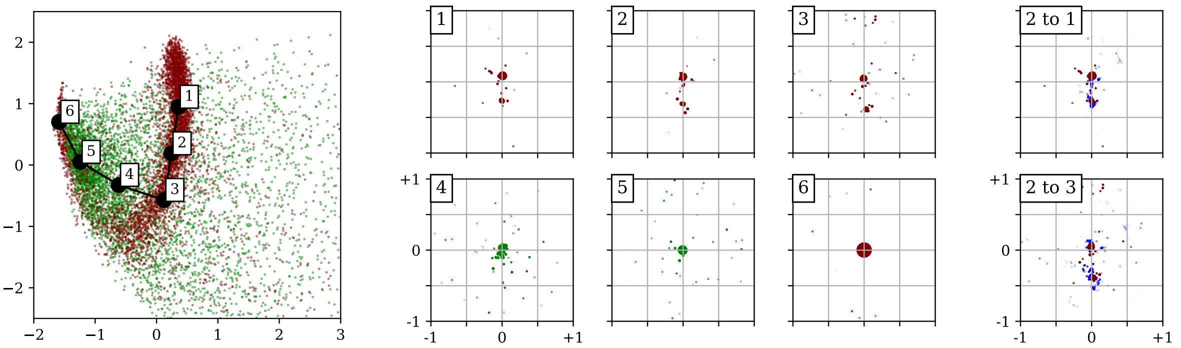


Figure 3.8: *Left*: Distribution of 10k W (red) and QCD (green) jets in the tangent space with respect to the first two dominant PCA modes for HK with $\kappa = 1$ with some positions in coefficient space marked. *Middle*: Jets in $(y - \phi)$-plane with first two PCA coefficients closest to the marks, color indicating jet type. From mark 1-3 the lower prong is moving and becoming weaker, while background noise increases. Mark 4 shows a single prong with strong noise. At mark 6 there is a distinct cluster of W jets with a single, strongly focused prong. *Right*: Approximating jets 1 and 3 (red) by the exponential map (blue) from jet 2, using the difference of the first two PCA coefficients as tangential direction. While this cannot account for the considerable variation that lies orthogonal to this 2D plane, it correctly describes the movement of the lower prong and the increase in peripheral noise. Figure copied from [34].

physically meaningful changes of the jets that can locally be approximated linearly via the exponential map. For other values of $\kappa \in [0.1, \infty]$, PCA yields qualitatively similar results.

Next, we apply linear discriminant analysis (LDA). Again, LDA assumes that samples from both classes are drawn from two Gaussian distributions with different means but identical covariance matrices. One then infers the hyperplane that optimally separates the two classes in a Bayesian sense. Let the hyperplane be parametrized by a unit normal vector t and a point z in the plane. On the one hand, LDA serves as a simple linear classifier where samples are labeled according to which side of the hyperplane they lie in, i.e., the predicted class of sample x_i depends on the sign of $\langle x_i - z, t \rangle$. On the other hand, we can analyze whether the direction t has a physical meaning and use LDA as a visualization tool.

The first row in Figure 3.9 shows the distribution of the projection coordinate $\langle x_i - z, t\rangle$ for W_2 and HK with various κ. We find that for $\kappa = 0.1$, HK best separates the two classes, which is later confirmed by the superior performance of the LDA classifier; see Table 3.2. Similar to PCA, applying the exponential map to the discriminating direction t relative to the sample mean do not yield physically valid results, presumably for the same reasons. Instead, we visualize the direction in another way: for some $\lambda \in \mathbb{R}$ we find the sample x_i such that $\langle x_i - z, t\rangle$ is closest to λ, i.e., among all samples x_i is closest to the hyperplane given by $\{x | \langle x - z, t\rangle = \lambda\}$. We vary λ on the order -3σ to 3σ, where σ denotes the standard variation of the samples along the direction t. This is shown in the lower two rows of Figure 3.9 for W_2 and $\mathrm{HK}_{\kappa=0.1}$, where we also record how many QCD and W jets there are in each λ bin.

For both W_2 and $\mathrm{HK}_{\kappa=0.1}$, the chosen jets transition from having a single mode to having two modes as λ increases, suggesting that the direction t successfully encodes the major topological difference between two-pronged W jets and more diffuse single-pronged QCD jets. A clearer separation is obtained for $\mathrm{HK}_{\kappa=0.1}$, whose class purity is slightly higher than W_2 in each λ bin.

3.4.2 Tagging Results and Discussions

We now consider kNN and SVM for the jet classification task. For kNN, we test k in $[10, 200]$ with an increment of 10, whereas for SVM we again test $11 \times 11 = 121$ pairs of C and γ both in $[10^{-5}, 10^{5}]$. A training set of 5000 jets, a validation set of 2500 jets and a test set of 2500 jets are used for both kNN and SVM to tune their hyperparameter(s). For LDA, the training and validation sets are instead merged.

Table 3.2 summarizes the results, including true positive rate (TPR) and false positive rate (FPR), for various κ values. The approximate run time on Google Colab is also given

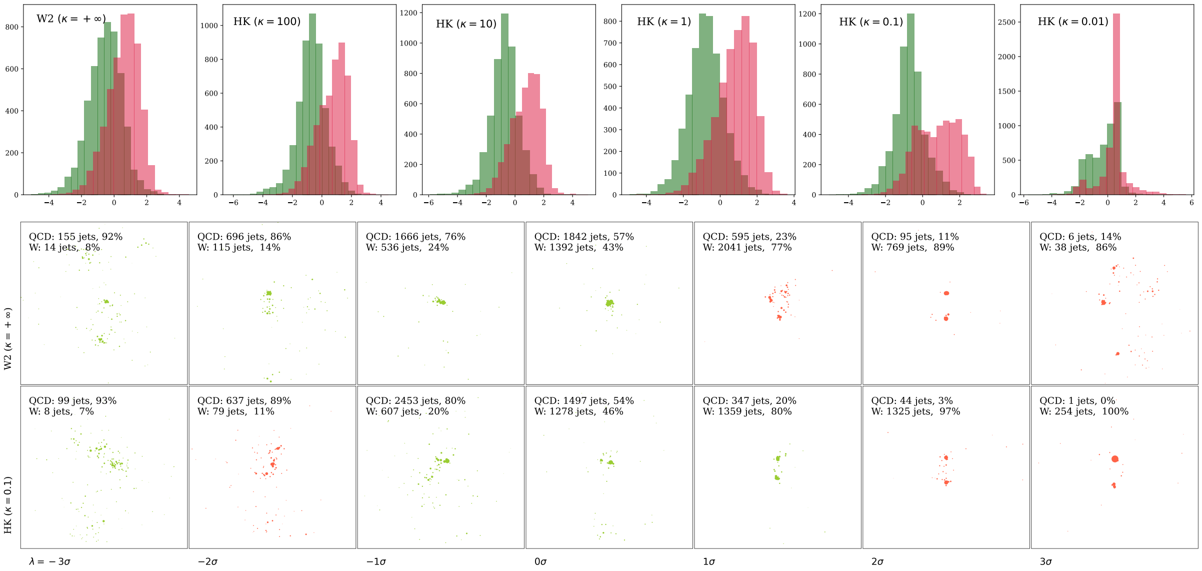


Figure 3.9: *Upper*: Histograms of the distribution of the LDA projection coordinate $\langle x_i - z, t\rangle$ with various OT distances. *Lower two rows*: Displays in the $y - \phi$ plane ($-1 \leq y \leq 1, -1 \leq \phi \leq 1$) of jets x_i such that $\langle x_i - z, t\rangle$ is closest to λ where $\lambda = -3$ to 3σ. In all plots, red denotes W jets and green is for QCD jets. Figure copied from [34].

to demonstrate the practicality of the linear framework. We see that AUC peaks when $\kappa = 0.1$ for LDA and 0.5 for both kNN and SVM, with a relative increase in performance of 10.2% for LDA, 3.4% for kNN, and 1.7% for SVM, with respect to the linear W_2 baseline. To gain a rough feeling of the performance gain, the AUC increase of using a large neural network over the optimal transport approach is only about 2%.

In addition, the gain of HK over W_2 is stronger on the relatively simplistic classifiers kNN and LDA and not as pronounced on the more sophisticated SVM classifier. This suggest that the W_2 representation does contain almost as much information about the jet class as the HK representations and sufficiently complex classifiers can extract it. In light of interpretability of classification results it may still be preferable to choose a representation where also simpler methods work well.

Compared to W_2, the HK distance requires the tuning of the parameter κ. The additional (computational) complexity of this step is however manageable, as a rough estimate for κ can be obtained from physical intuition. Usually, κ should be on a par with the length scale present in the dataset. In our case, it is the typical separation between particles in a jet, which is on the order of 0.1. This matches nicely with our observation of the optimal κ values for the present classification task.

The first-pass estimate of κ can subsequently be refined by cross validation. Table 3.2 indicates that the classification behavior is relatively robust with respect to κ over almost one order of magnitude. Therefore, a coarse cross validation parameter search is sufficient and too much fine-tuning is unnecessary.

In order to understand the fluctuation of AUC quantitatively, we repeat the analysis (without $\kappa = 0.7, 0.3$) on two additional datasets, each again containing 10k W and QCD jets simulated in the same way. As hoped, we observe that the deviation in AUC is no larger than 10% and the general trend is the same for all datasets; see Figure 3.10.

Moreover, we investigate the impact of the reference measure on classification performance. We include a different reference which is the linear mean of all QCD jets in the dataset after rasterization to a grid, shown in the upper row of Figure 3.10. We call these new measures the QCD references. The three datasets are then analyzed using their respective QCD references.

The AUC curves are shown in Figure 3.10. We observe that in general an adapted choice of the reference measure slightly improves model performance in the best κ range, yet the performance deteriorates faster when $\kappa \to 0.01$ compared to the uniform measure. On the whole, the classification performance using either reference measure is comparable to each other.

Table 3.2: Results for the W *vs.* QCD jet tagging task using LDA, kNN and SVM on the LOT embeddings for various length scale parameters κ ($\kappa = +\infty$ denotes balanced W_2).

length scale κ		$+\infty$	100	10	5	1	0.7	0.5	0.3	0.1	0.05	0.01
LDA	**AUC**	**0.694**	**0.733**	**0.746**	**0.747**	**0.752**	**0.751**	**0.748**	**0.760**	**0.765**	**0.763**	**0.642**
	TPR	0.684	0.684	0.703	0.721	0.724	0.740	0.736	0.692	0.704	0.731	0.770
	FPR	0.296	0.218	0.211	0.226	0.220	0.239	0.239	0.171	0.174	0.205	0.486
	run time	several seconds										
kNN	**AUC**	**0.821**	**0.818**	**0.819**	**0.818**	**0.829**	**0.841**	**0.849**	**0.847**	**0.821**	**0.772**	**0.671**
	TPR	0.771	0.763	0.768	0.763	0.760	0.791	0.798	0.809	0.821	0.783	0.733
	FPR	0.128	0.127	0.130	0.126	0.102	0.110	0.100	0.114	0.181	0.238	0.390
	hyperpar. k	30	20	30	20	10	20	10	20	10	10	30
	run time	1.5 hours										
SVM	**AUC**	**0.842**	**0.842**	**0.842**	**0.841**	**0.849**	**0.851**	**0.856**	**0.853**	**0.845**	**0.806**	**0.694**
	TPR	0.817	0.819	0.817	0.819	0.823	0.829	0.832	0.829	0.788	0.741	0.787
	FPR	0.133	0.134	0.134	0.137	0.126	0.127	0.120	0.124	0.099	0.128	0.401
	hyperpar. C	1	1	1	1	1	1	1	1	1	10	10
	hyperpar. γ	100	100	100	100	100	100	100	100	1000	1000	100000
	run time	5 hours										

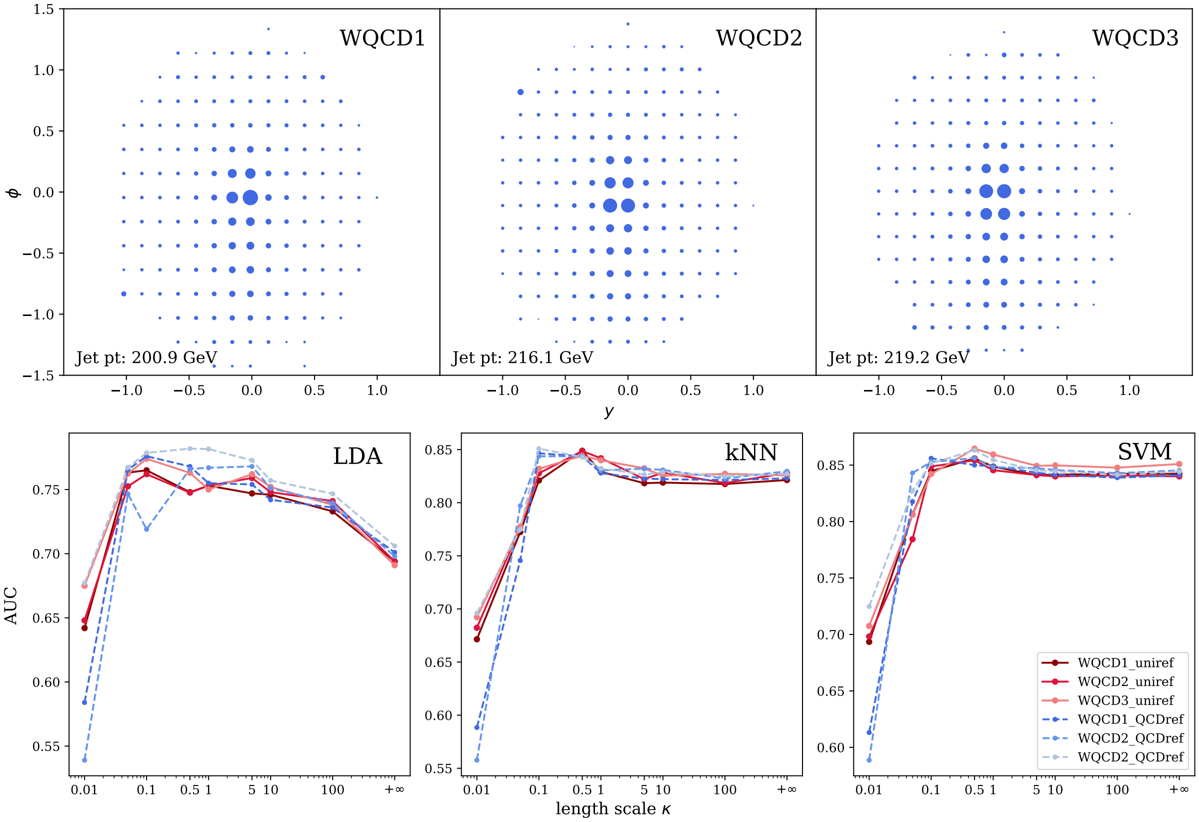


Figure 3.10: *Upper*: QCD reference jets in the $y - \phi$ plane for the three W *vs.* QCD datasets, obtained by averaging all QCD jets in the respective dataset. Note that WQCD1 is the same dataset used in Table 3.2. *Lower*: AUC scores for LDA, kNN and SVM on the three datasets with W_2 and HK metrics of various κ's. Red solid lines show the results using the uniform reference measure, whereas blue dashed lines are obtained using the QCD reference measures. Figure copied from [34].

In conclusion, our study suggests that allowing mass to be generated and annihilated rather than only transported have a positive effect on picking out W jets from QCD background. In the next section, we present a more detailed analysis on the comparison of the balanced W_2 and unbalanced HK distances for jet tagging, especially for jets with wider total p_T differences.

3.5 Balanced *vs.* Unbalanced OT: Which Metric to Use?

We now tackle the next question—which is the best metric for the space of collider events? In this section, we study the performance of the W_2 and HK distances as a function of the scale parameter κ, jet p_T range, and the choice of reference measure. Note again that we use a slightly different notation than the original paper [36]. Here LOT includes both linearized W_2 and linearized HK approximations. The use of the original acronym PLUOT is discontinued, in favor of the new name LinHK for increased clarity and consistency within the thesis.

As before, we focus on the task of distinguishing boosted W jets and QCD background jets. We consider simulated data consisting of 200k W jets and QCD jets, generated as in Section 3.3. Proton-proton collision events at $\sqrt{s} = 14$ TeV are simulated in MADGRAPH 2.9.2 [99] with W bosons being pair produced, gluons generated via $q\bar{q} \to Z \to \nu\bar{\nu}g$, and quarks via $qg \to Z \to \nu\bar{\nu}q$. The particles are then hadronized and decayed in PYTHIA 8.302 [100], where default tuning and showering parameters are used. Afterwards, we cluster the events into jets using FASTJET 3.3.4 [84] with anti-k_T algorithm (jet radius $R = 1$), where at most two jets are kept with $|y| \leq 1.7$ and $|\phi| \leq \frac{\pi}{2}$.

Before calculating their LOT embedding, we boost and rotate the jets to center the jet 4-momentum and vertically align the principal component of the constituent p_T flow in the $y - \phi$ plane using the `EnergyFlow` package, and normalize all jets to have unit total p_T; the pre-processing is the same as in Section 3.3. For LinHK_κ, we consider values of κ ranging over several orders of magnitude, i.e., $\kappa \in [0.01, 100]$.

Once the Euclidean embedding for each jet is acquired via the linearization scheme, we again employ kNN and SVM for classification. We consider the datasets consisting of either 10k or 200k W and QCD jets. For datasets with 10k jets, we use 5000 jets to train kNN and SVM, 2500 for validation in order to pick the best model hyper-parameter(s), and the remaining 2500 jets as the test dataset to obtain the model performance. We try $k \in [10, 100]$ with an increment of 10 for kNN, and $C, \gamma \in [10^{-2}, 10^{5}]$ for SVM where only powers of 10 are considered (the ranges are smaller than before based on our enhanced experience with the possible best values). When dealing with the full 200k dataset, we use 150k jets to train the models and 50k to evaluate the performance, where the model hyper-parameter(s) are already picked by smaller runs with the 10k datasets.

We compare the tagging performance of kNN based on the the LOT framework to that of N-subjettiness ratio $\tau_{21} = \tau_2/\tau_1$, where τ_N is determined using the `Nsubjettiness` plug-in package in FastJet [85, 102]. Another benchmark is the pairwise EMD distance matrix [4] coupled with the same machine learning models. We test the EMD both on normalized jets, where the jets are first rescaled to have $p_T = 1$, as well as on unnormalized jets using the modified EMD*. The ability to compare both normalized and unnormalized jets is implemented by a built-in function in the `EnergyFlow` package, with the parameters $R, \beta = 1$ and the normalization parameter `norm` set respectively to `True` or `False`.

In a similar manner, the LinHK framework also presents us with two options to calculate the Euclidean embedding. One way is to compute the unbalanced HK distances directly between jets with different total p_T. Alternatively, we can again normalize the jets so that each has $p_T = 1$ and then compute the unbalanced HK$_\kappa$ distance between the normalized jets. We emphasize that, even when two jets have equal total p_T, as in the case of balanced OT, the HK distance still allows for local mass to be created and destroyed.

As the normalized and unnormalized approaches to HK are related by simple scaling transformations, in practice we begin by computing the Euclidean embedding of normalized jets and then recover the embedding of the unnormalized jets. Hereafter, we abbreviate the distances calculated on normalized jets with a subscript of N and those obtained for unnormalized jets with subscript *unN*.

3.5.1 Widen the p_T Range

Our previous studies of jet classification based on OT have been relatively insensitive to differences in total jet p_T, typically considering events drawn from narrow (50 GeV) p_T bins. Indeed, in Section 3.3, it was observed that classification based on balanced optimal transport distances between normalized jets drawn from a 50 GeV p_T bin modestly outperformed unbalanced optimal transport distances using the modified EMD formulation.

To better assess the effects of unbalanced samples, we explore jets drawn from a broader range of total p_T, extending from $[500, 550]$ GeV to $[500, 1500]$ GeV. This is achieved by stacking 20 datasets, each containing 10k jets with a p_T bin of 50 GeV, i.e., $p_T \in [500, 550]$ GeV for the first dataset, $p_T \in [550, 600]$ GeV for the second, and so forth. In this way, in addition to the 20 datasets each with 50 GeV p_T bin width, we have a combined dataset of 200k jets in which the total jet p_T is approximately uniformly distributed between 500 and 1500 GeV. We can now study the classification performance as a function of the p_T range of the simulated events, comparing the tagging performance of W *vs.* QCD jets whose total $p_T \in [500, 550]$ GeV or $[500, 1500]$ GeV.

The three OT distances examined here include: 1) the EMD distance on normalized jets (EMD_N) and its modified version on unnormalized jets (${\text{EMD}^*}_{unN}$); 2) the balanced W_2 distance on normalized jets; and 3) the HK distance on both normalized and unnormalized jets (denoted as HK^κ_N or HK^κ_{unN}). The N-subjettiness ratio τ_{21} is also computed for each jet as a benchmark. For the HK^κ distance, we consider the κ values $+\infty$, 100, 10, 1, 0.5, 0.4, 0.3, 0.2, 0.1, 0.07, 0.05, 0.03, 0.01, with $\kappa = +\infty$ denoting the W_2 distance.

Here the reference measure is taken to be a uniform jet with a total $p_T = 750$ GeV

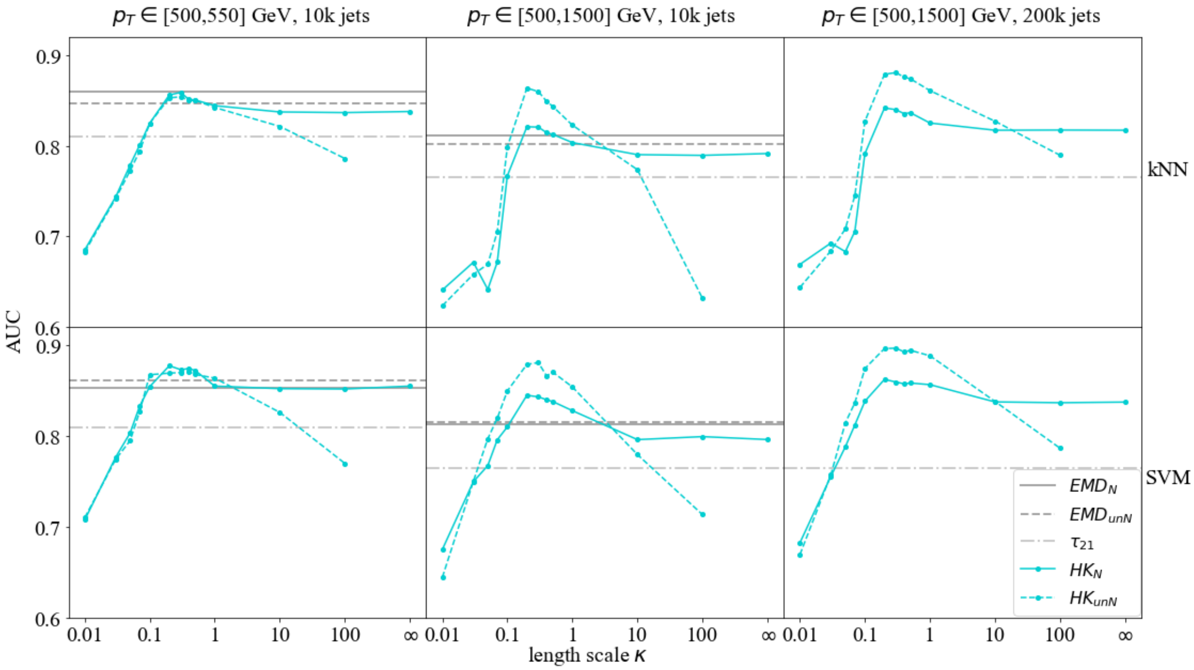


Figure 3.11: AUC scores for classifying W *vs.* QCD jets using kNN and SVM models coupled to linear W$_2$/HK$^\kappa$ embedding with $\kappa \in [0.01, +\infty]$. Jet p_T ranges from 500 to 550 (1500) GeV in the first (second, third) column. The datasets for Column 1 and Column 2 (Column 3) have 10k (200k) jets. Solid (dashed) blue lines show the results calculated on normalized (unnormalized) jets; horizontal gray solid (dashed) lines use the EMD metrics on normalized (unnormalized) jets; and grey dash-dotted lines give the performance using τ_{21} as the discriminator.

and $15 \times 15 = 225$ particles. Since it is impossible to calculate and store the entire distance matrix for 200k jets using the EMD approach with reasonable computational resources, we only compute EMD distances on the 10k datasets, whereas the linear W$_2$ and HK embedding can be calculated efficiently for the full 200k datasets.

Figure 3.11 shows the tagging performance in terms of the AUC score. A discussion of the general trends of the tagging performance not specific to the present task is deferred to the later *Discussions* section, as well as a more detailed table where results from other tasks are also included.

As can be seen in Figure 3.11, for jets drawn from a 50 GeV-wide p_T bin (column 1), classification performance on either normalized or unnormalized jets is almost indistinguishable for LinHK with small κ values ($\kappa \leq 1$). The EMD approach also produces similar AUC scores regardless of whether or not the jets are normalized, with kNN slightly preferring the normalized approach and SVM favoring the unnormalized version. The percentage differences in the AUC are within 1.5%, consistent with statistical fluctuations. Such behavior is to be expected since normalization should not make a big difference when the total p_T difference among jets is small. Additionally, the tagging performance of the LOT approximation, including W_2 and HK (with the exception of $\mathrm{HK}^{\kappa}_{unN}$ for large κ) approaches the same (or better, in the case of SVM) level of accuracy of the EMD method, with far less computational expense.

However, the effect of normalization becomes significant when the p_T bin width is broadened. For jets with $p_T \in [500, 1500]$ GeV (10k for column 2 and 200k for column 3), the HK distance with κ in its optimal range calculated directly on the unnormalized jets (dashed blue lines) gives superior performance to the normalized jets (solid blue lines), whether we use kNN or SVM as the coupled model. The increase in AUC reaches about 5% at its peak when $\kappa \sim 0.2$. There the AUC from the HK distance, whether normalized or not, is noticeably higher than when using the EMD distance.

Interestingly, such performance gain is not observed in the EMD approach. Here it makes no notable difference whether we use EMD_N (solid gray line) or $\mathrm{EMD}^*{}_{unN}$ (dashed gray line). This implies that though the difference in total jet p_T has potential discriminating power, not all approaches to unbalanced optimal transport take advantage of it. A simple difference term like $|p_T(\text{jet } 1) - p_T(\text{jet } 2)|$, as included in the modified EMD formulation, does not lead to improved discrimination for samples drawn from a larger p_T range. In contrast, unbalanced HK, especially HK_{unN}, appears to take better advantage of this information by allowing *local* mass to be created and destroyed in addition to being transported.

Note that, while the original formulation of the EMD in the particle physics literature considered a fixed scale parameter $R = \frac{\kappa}{2} \geq \max_{ij} d_{ij}/2$ coinciding with the jet clustering radius, one could perform a similar analysis by using the more general partial transport distance, investigating how different choices of $R = \frac{\kappa}{2}$ lead to different amounts of creation and destruction and, potentially, improved AUC in certain regimes. However, due to the fact that such metrics lack a Riemannian structure amenable to linearization, the analysis of finding the optimal parameter $R = 2\kappa$ would be extremely computationally intensive.

3.5.2 Vary the Reference Measures

In the LOT framework, we are in principle free to pick any reasonable measure as our reference jet. Ideally, the choice of a reference measure should not exert too large an impact on the calculated linear W_2/HK embedding and the downstream tagging performance. As a first study, we examine the effect of varying the resolution of the reference measures on classification.

We choose five uniform reference jets consisting of 4×4, 8×8, 15×15 (the default), 30×30, and 60×60 particles, respectively denoted as "uniref4", "uniref8", "uniref15", "uniref30" and "uniref60". All reference jets have a total $p_T = 750$ GeV, distributed uniformly on the $y - \phi$ rectangle $[-1.7, 1.7] \times [-\pi/2, \pi/2]$. The default "uniref15" has about the same number of particles as in a typical W or QCD jet in our sample of simulated events. The inter-particle spacing l of these reference jets differs widely, ranging from roughly 0.05 to 0.85. This defines yet another length scale in addition to the HK scale parameter κ, the jet clustering radius R, and the characteristic angular separation of the partonic decay products of a boosted particle of mass m, which is proportional to m/p_T. It is natural to consider the interplay between all these length scales in determining what constitutes a reasonable measure in practice.

In Figure 3.12, we show the distribution of the Euclidean norms of the LOT coor-

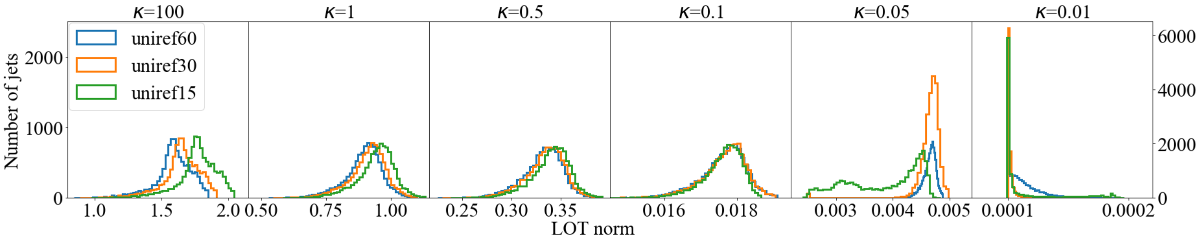


Figure 3.12: Distribution of the LOT norm, i.e., the distance from each jet's LOT coordinate to the origin, for the HK distance with $\kappa = 100, 1, 0.5, 0.1, 0.05, 0.01$. The uniform reference measures used include "uniref15" with 15×15 particles, and similarly "uniref30" and "uniref60". The y coordinate of the rightmost plot follows the scale on the right, while the other plots follow the scale on the left.

dinates of 10k jets ($p_T \in [500, 550]$ GeV) with HK^κ using "uniref15", "uniref30", and "uniref60". [2] As κ is decreased from large values $\kappa \sim 100$ the distribution of the norms using the HK^κ distance becomes more and more similar for different reference measures. The closest agreement occurs for $\kappa \sim 0.1$, which we will see later is the κ value that gives the optimal tagging performance. As κ is decreased below $\kappa \sim 0.1$ and we enter a scaled Euclidean image difference regime, the discrepancy of the norms using different reference measures becomes noticeable again. We will see that this instability with respect to the chosen reference measure translates to deterioration of the tagging performance for small κ values.

Figure 3.13 shows the tagging performance on 10k jets with total $p_T \in [500, 550]$ GeV (first row) and $[500, 1500]$ GeV (second row) using EMD_N, $\mathrm{EMD}^*{}_{unN}$; HK_N, HK_{unN}; and the N-subjettiness ratio τ_{21}. Tagging performance is plotted in terms of AUC as a function of κ for the HK distances.

Apart from similar behaviors already discussed above, we observe here that the peak tagging performance is roughly the same for all reference measures except "uniref4", which does not attain tagging performance comparable to any EMD distance for any

[2] As we will see, the "uniref4" and "uniref8" reference measures are too coarse to capture the relevant structure of the jets for any value of κ, and the distribution of Euclidean norms for these measures are correspondingly omitted from Figure 3.12.

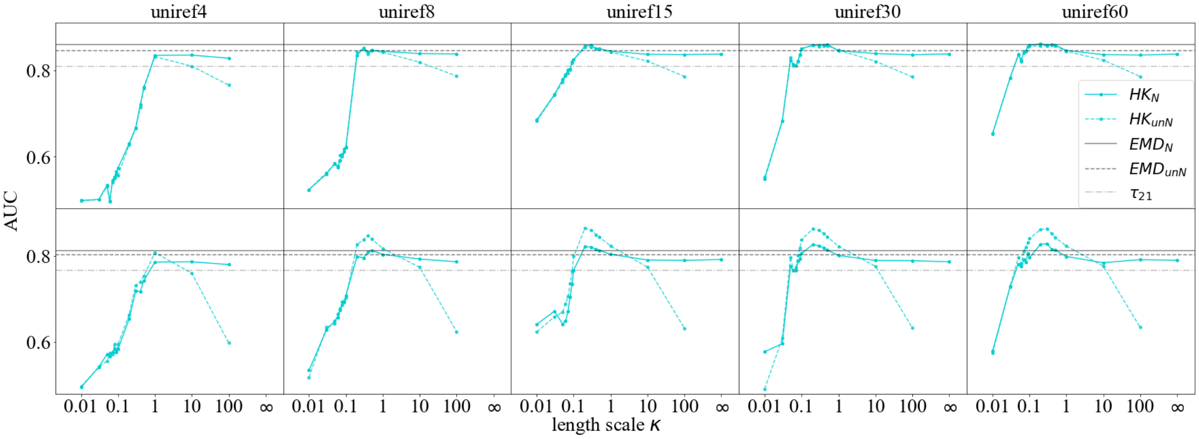


Figure 3.13: AUC scores for classifying 10k W *vs.* QCD jets using different reference measures, with “uniref4”, “uniref8”, “uniref15”, “uniref30”, and “uniref60” (from left to right). The machine learning model used here is kNN. Jet p_T is in between 500 and 550 (1500) GeV in the first (second) row. Solid (dashed) blue lines show the results calculated on normalized (unnormalized) jets for W$_2$/HK distance; horizontal grey solid (dashed) lines use the EMD metrics on normalized (unnormalized) jets; and grey dash-dotted lines give the tagging performance of τ_{21}.

value of κ. Although “uniref8” yields tagging performance comparable to EMD$^*_{unN}$ for jets with total $p_T \in [500, 550]$ GeV, it does not reach the tagging performance of EMD$_N$. In contrast, the tagging performance of LinHK using “uniref15”, “uniref30”, and “uniref60” meets or exceeds the tagging performance of the EMD distances for optimized values of κ. This suggests that the classification performance of the linear W$_2$/HK distance is rather robust to the choice of the reference for “uniref15” and finer measures. Considering that the finest reference measure under consideration (“uniref60”) incurs a relatively high computational cost without significant improvement in tagging performance, in what follows we largely favor the default 15 × 15 reference jet, reserving some comparisons with “uniref30” for the table in the later *Discussions*.

Table 3.3 lists the κ value that produces the best AUC score for each task using HK$_N$ and HK$_{unN}$ metrics. Ignoring “uniref4”, the optimal value κ_{best} lies between 0.2 and 0.5 for all others, regardless of the inter-particle spacing l. No obvious relationship is observed between l and κ_{best}.

Table 3.3: Optimal κ values and their corresponding AUC scores for kNN classification of W *vs.* QCD jets using different reference measures.

Jet p_T (GeV)		**[500, 550]**		**[500, 1500]**	
Reference		**HK$_{\mathbf{N}}$**	**HK$_{\mathbf{unN}}$**	**HK$_{\mathbf{N}}$**	**HK$_{\mathbf{unN}}$**
uniref4	κ_{best}	10	1	10	1
(4×4)	AUC	0.835	0.832	0.786	0.807
uniref8	κ_{best}	0.3	0.3	0.5	0.4
	AUC	0.852	0.849	0.813	0.847
uniref15	κ_{best}	0.3	0.3	0.2	0.2
	AUC	0.859	0.854	0.821	0.863
uniref30	κ_{best}	0.5	0.2	0.2	0.2
	AUC	0.860	0.859	0.826	0.862
uniref60	κ_{best}	0.2	0.4	0.3	0.3
	AUC	0.862	0.858	0.828	0.863

3.5.3 Discussions

Table 3.4 presents the detailed results for selected tasks considered above using the kNN model. On average, only about an hour is needed to calculate the LOT distance for 10k jets on a laptop, with further speedup in the LinHK embedding for smaller κ values. In contrast, computing EMD for 10k jets takes approximately 15 hours for jets drawn from a 50 GeV-wide p_T bin, and 30-40 hours for jets drawn from a 1 TeV-wide p_T bin.

The tagging performance of HK as shown in Figures 3.11 and 3.13 exhibits three distinct regimes as a function of κ. In the regime where mass-creation/destruction dominates ($\kappa \lesssim 0.1$), the AUC scores for both HK$_N$ and HK$_{unN}$ are comparable and decrease with decreasing κ. Since no mass is allowed to be moved a distance more than $\frac{\pi}{2}\kappa$, when κ becomes so small such that $\frac{\pi}{2}\kappa < l$ (where l is the inter-particle spacing of the reference jet), mass transportation is largely forbidden when computing the distance between a jet and the reference measure. Furthermore, in this regime the assumption that, for each particle $\tilde{x}_j$ in the jet, there exists a particle x_i in the reference measure so that $\|x_i - \tilde{x}_j\| < \frac{\kappa\pi}{2}$ is often violated, causing the linearization itself to break down. While this breakdown could be avoided in a continuum formulation of the linearization, one would still have to contend with the fact that, as $\kappa \to 0$, the rescaled Hellinger-Kantorovich metric converges to the Hellinger metric, in which all information on the spatial distribution of the jets is discarded and their distance is based purely on the difference between their energies at each location. We observe this breakdown at the level of the AUCs in Figure 3.12, considering the value $\kappa = 0.01$.

At the other end, at large κ, the tagging performance using HK_N stabilizes for $\kappa \gtrsim 1$, whereas the AUC score deteriorates significantly using HK_{unN}. As κ grows sufficiently large, it becomes increasingly expensive to create or destroy mass. Once we enter this transport-only regime, κ no longer plays any role for HK_N. On the other hand, whenever the total energies of the events are unequal, HK_{unN} diverges to $+\infty$ as $\kappa \to +\infty$.

In between these two extremes, $0.1 \lesssim \kappa \lesssim 1$, the tagging performance of both HK_N and HK_{unN} is optimized, matching or exceeding the EMD approach. In this regime both mass transportation and creation/destruction are relevant. Unfortunately, we have not observed any strong correlations between the optimal value κ_{best}, reference spacing l, the jet clustering radius R, and the typical angular separation of boosted partonic decay products $\propto m/p_T$, and no definite conclusion can be drawn at this stage regarding the dependence of κ_{best} on various jet length scales. We leave this question to future studies.

3.6 Pileup Robustness of the OT Framework

Radiation emitted from secondary collisions, commonly known as *pileup* (PU), may overlap with that of the primary interaction of interest, posing a significant challenge to the extraction of valuable insights from data collected at hadron colliders. A rough estimation states that each pileup vertex typically adds about 600 MeV of energy per unit rapidity per unit azimuth [103, 104], making up as high as 30% of energy density coming

Table 3.4: AUC scores for kNN classification of W *vs.* QCD jets.

Ref Jet	κ	Normalization	p_T		
			$[500, 550]$ GeV, 10k jets	$[500, 1500]$ GeV, 10k jets	$[500, 1500]$ GeV, 200k jets
uniref15	$+\infty$	N	0.838	0.791	0.817
	100	N	0.836	0.789	0.817
		unN	0.786	0.632	0.790
	10	N	0.837	0.790	0.817
		unN	0.821	0.774	0.827
	1	N	0.844	0.803	0.825
		unN	0.842	0.823	0.861
	0.5	N	0.850	0.812	0.836
		unN	0.850	0.843	0.874
	0.2	N	0.856	0.821	0.842
		unN	0.853	0.863	0.879
	0.1	N	0.825	0.767	0.791
		unN	0.825	0.799	0.827
	0.05	N	0.779	0.642	0.683
		unN	0.773	0.669	0.708
	0.01	N	0.685	0.641	0.669
		unN	0.683	0.624	0.644
uniref30	$+\infty$	N	0.838	0.786	0.815
	100	N	0.836	0.789	0.815
		unN	0.785	0.633	0.791
	10	N	0.839	0.789	0.815
		unN	0.821	0.776	0.827
	1	N	0.846	0.801	0.827
		unN	0.847	0.822	0.860
	0.5	N	0.860	0.813	0.839
		unN	0.856	0.844	0.874
	0.2	N	0.857	0.826	0.842
		unN	0.859	0.862	0.880
	0.1	N	0.851	0.806	0.827
		unN	0.849	0.837	0.861
	0.05	N	0.823	0.775	0.802
		unN	0.830	0.797	0.825
	0.01	N	0.549	0.577	0.566
		unN	0.552	0.492	0.567
EMD	N		0.859	0.812	N/A
	unN		0.846	0.802	
τ_{21}			0.810	0.766	0.765

from the primary collision event. The previous data taken by the ATLAS and CMS collaborations at the LHC contain on average 20 pileup events per bunch crossing, i.e., $\langle N_{PU} \rangle \sim 20$, which has now increased to $\langle N_{PU} \rangle \sim 80$ in the current Run 3 experiments. The planned upgrades to future High Luminosity LHC (HL-LHC) will only make things worse, as it is expected to have as high as $\langle N_{PU} \rangle \sim 200$ for Run 4-5.

Such radiation contamination from pileup significantly reduces the efficacy of many commonly used jet physics observables [105, 106, 107], such as jet mass and dijet mass, where in [108] the impact of different levels of pileup on dijet mass is studied. This in turn motivates the invention of various pileup mitigation techniques [109, 110, 111, 112, 113, 114]. Pileup mitigation has recently been recast in the language of optimal transport [13], but the robustness of OT-based approaches to jet classification has yet to be studied. Here we present a first study of the robustness of OT-based approaches in the presence of pileup (or any other form of uniform noise).

Again we use the same W and QCD dataset with $p_T \in [500, 550]$ GeV and jet radius $R = 1$ as before. We now need to add in contamination, where we generate the so-called pileup templates in Pythia. The actual number of pileup events per bunch crossing follows a Poisson distribution around $\langle N_{PU} \rangle$. We consider three different pileup templates with $\langle N_{PU} \rangle = 20, 80, 140$, according to the experimental benchmarks. These pileup templates are then added to each event and FastJet is used to group the pileup-contaminated events into jets. We then follow the same procedure as before, applying the LOT framework to the pileup-contaminated jets.

We include three reference measures: the default 15×15 uniform reference; the 30×30 uniform reference; and an additional "pileup" reference jet (termed as "PUref") picked from one of the pileup templates for each value of $\langle N_{PU} \rangle$. For example, when examining jets contaminated by pileup with $\langle N_{PU} \rangle = 80$, the reference measure is taken to be another Poisson distribution with $\langle N_{PU} \rangle = 80$. The motivation behind the choice of “uniref30” is that since the number of particles in the reference is close to that of the jets contaminated by pileup with $\langle N_{PU} \rangle = 80, 140$, “uniref30” should better capture the true underlying differences between W *vs.* QCD jets not obscured by the superficial pileup addition.

Again, the N-subjettiness ratio τ_{21} serves as a benchmark, where τ_{21} is computed both on the datasets with and without pileup. The one without pileup is generated by pruning the contaminated datasets, accomplished in FastJet by a pruner that reclusters the jets with Cambridge-Aachen algorithm and removes constituent particles that are soft or at large angles with other particles [115, 116].

Figure 3.14 displays the resulting AUC *vs.* κ curves, where we use kNN coupled with LinW$_2$ or LinHK with $\kappa = +\infty, 10, 1, 0.5, 0.2, 0.1,$ and 0.05 on both normalized and unnormalized jets. It is clear from the figure that comparing to τ_{21} (horizontal lines), the tagging performance of LOT behaves rather well and does not decay significantly as pileup increases. Especially for high pileup scenarios, the AUC scores of kNN+W$_2$/HK distances on un-pruned jet samples using any of the three references are far better than the corresponding AUCs of kNN+τ_{21}. For $\langle N_{PU} \rangle = 140$, τ_{21} on pruned jets behaves much worse than that on un-pruned jets, corroborating the observation in [117] that N-subjettiness on groomed jets is less discriminant than being computed on ungroomed jets. Table 3.5 summarizes the AUC results for the LOT framework using the three different reference jets, as well as those for τ_{21}.

More studies need to be performed in order to examine in detail the influence of background contamination such as pileup on OT-based metrics, but its potential advantage over traditional methods is already clear from our preliminary study.

Table 3.5: AUC scores for kNN classification of W *vs.* QCD jets with different levels of pileup.

Ref Jet	κ	Normalization	$\langle N_{PU} \rangle = 20$	$\langle N_{PU} \rangle = 80$	$\langle N_{PU} \rangle = 140$
uniref15	$+\infty$	N	0.801	0.768	0.754
	10	N	0.802	0.770	0.750
		unN	0.777	0.752	0.736
	1	N	0.820	0.790	0.767
		unN	0.818	0.794	0.767
	0.5	N	0.821	0.796	0.773
		unN	0.831	0.796	0.774
	0.2	N	0.830	0.787	0.766
		unN	0.829	0.787	0.767
	0.1	N	0.812	0.775	0.739
		unN	0.812	0.779	0.743
	0.05	N	0.740	0.714	0.686
		unN	0.739	0.714	0.682
uniref30	$+\infty$	N	0.799	0.772	0.757
	10	N	0.799	0.775	0.759
		unN	0.775	0.756	0.743
	1	N	0.819	0.786	0.765
		unN	0.819	0.792	0.761
	0.5	N	0.824	0.794	0.770
		unN	0.819	0.795	0.768
	0.2	N	0.835	0.796	0.766
		unN	0.830	0.798	0.765
	0.1	N	0.832	0.798	0.752
		unN	0.832	0.800	0.757
	0.05	N	0.808	0.771	0.730
		unN	0.813	0.769	0.727
PUref	$+\infty$	N	0.797	0.777	0.753
	10	N	0.795	0.776	0.752
		unN	0.779	0.758	0.734
	1	N	0.805	0.785	0.756
		unN	0.801	0.788	0.754
	0.5	N	0.823	0.792	0.770
		unN	0.822	0.790	0.770
	0.2	N	0.838	0.800	0.769
		unN	0.832	0.799	0.770
	0.1	N	0.825	0.795	0.756
		unN	0.822	0.795	0.754
	0.05	N	0.787	0.768	0.725
		unN	0.781	0.772	0.720
τ_{21}			0.761	0.697	0.657
pruned τ_{21}			0.784	0.708	0.585

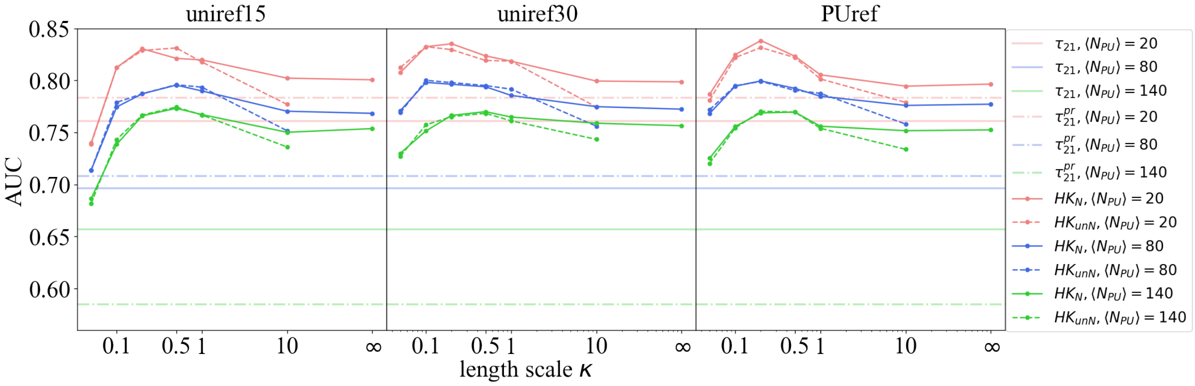


Figure 3.14: AUC scores for using kNN to classify 10k W *vs.* QCD jets with different amount of pileup where the average numbers of pileup particles in each event are $\langle N_{PU} \rangle = 20$ (red), 80 (blue), and 140 (green). From left to right, the reference measures used are the 15×15 uniform reference, the 30×30 uniform reference, and a jet drawn from the pileup template corresponding to each N_{PU}. As usual, solid (dashed) lines show the AUC scores using the LinW$_2$ or LinHK on normalized (unnormalized) jets, and solid horizontal lines give the tagging performance of τ_{21} on unpruned jets whereas the dash-dotted lines are the results using τ_{21} on pruned jets (denoted by τ_{21}^{pr}).

3.7 Upgrading the Optimal Transport Framework

Up until now, we have focused exclusively on the task of jet tagging. To this end, we have introduced and adopted the tools of balanced and unbalanced optimal transport, as well as developing their linearization scheme to improve practical utility. Of course, there is much more in collider physics than just jet tagging. One obvious next step is to move to the full LHC event level and consider event classification.

As a first try, one may treat an entire LHC event as if it were a single jet and directly apply the above OT methods on events without modification to the framework. We have attempted such a naïve implementation on a test dataset of W *vs.* QCD dijet events and observed that the classification performance suffers as the AUC drops below 0.65. A more careful examination suggests that there are at least three major complications arising in the case of event classification.

Event preprocessing is ambiguous: OT with Invariances

First, even before optimal transport can be applied, we need a way to preprocess the raw data so as to get rid of any superfluous difference in the energy flows, such as translations and rotations which are the invariances of the collider system. In the case of jets, preprocessing is not a large concern, as there is a well-defined procedure to center and rotate jets which has also been used throughout the previous sections. However, no such consensus exists for events and every developer needs to propose their own preprocessing scheme suitable to the particular framework they use. This makes it difficult to efficiently compare different frameworks, partially explaining the low number of event classification studies relative to jet tagging.

More importantly, there is no intrinsic preference of one preprocessing scheme over another. It is dubious how to rotate a full event on the $y-\phi$ plane; such a rotation even lacks physical meaning. Every preprocessing scheme is therefore equally *ad-hoc* and only serves the purpose of the particular statistical framework under consideration.

One way to bypass the preprocessing issue is if the framework itself is invariant under certain transformations of the energy flow. For example, people having been using techniques in geometric deep learning which preserve certain symmetries and invariances of the underlying data. In the case of optimal transport, ideally we would also want to make the distance itself invariant under the symmetries of the system. We term this new framework “OT with Invariance”.

The key idea here is to devise an appropriate reference measure which is invariant under the desired symmetries. Currently, we are focusing only on the 1D rotational symmetry in the $y-\phi$ plane. In this case, obviously, the discrete reference would consist of concentric rings with equally-angular-spaced dots. One then computes the LOT coordinate of each event with respect to this rotationally symmetric reference, which by design would give the same coordinate even if we rotate the original event (since it is equivalent to counter-rotating the reference).

Of course, there is still the possibility of an angular offset when comparing the LOT coordinates of two different events and calculating their Euclidean distance. But this can be resolved by scanning over all possible rotational angles (given by the angular resolution of the discrete reference) and picking the minimum distance. At least for the simple 1D rotation, the later step is straightforward and rather efficient computationally, since for each angle only the Euclidean distance calculation is required between each pair of the LOT coordinates.

We have implemented the above method and verified that it works as intended on toy data; see Figure 3.15. Here, the reference $\mathcal{R}$ consists of 10 concentric rings with an angular resolution of 6 degrees. The distributions being compared are two randomly sampled letter "S" in the normal upright position, which gives a $\mathrm{LinW}_{2,\mathcal{R}} = 0.174$ between them without any rotation. We then rotate one S by 43°, resulting in the green distribution S′. The $\mathrm{LinW}_{2,\mathcal{R}}$ pseudo-distance between S and S′ is then calculated every 6 degrees and correspondingly plotted on the right as a function of the rotational angle θ. Two valleys are clearly visible, with the lowest one giving a minimum distance of 0.166 obtained when $\theta = 42^{\circ}$, which is the closest angle to the actual 43° given the resolution of the reference. It is not surprising that the minimum distance obtained through the new method of OT with Invariance may be even lower than that of two upright S's ($0.166 < 0.174$), since the two S distributions are not exactly the same. Furthermore, the second minimum occurs when $\theta = 222^{\circ} = 42^{\circ} + 180^{\circ}$, exactly reflecting the 180° rotational symmetry of the letter "S".

Of course, this simple method would not work as well for high-dimensional rotations. One would then need a gradient descent or other optimization methods than the current brute-force exhaustive search over all possible values of the discrete 1D rotational

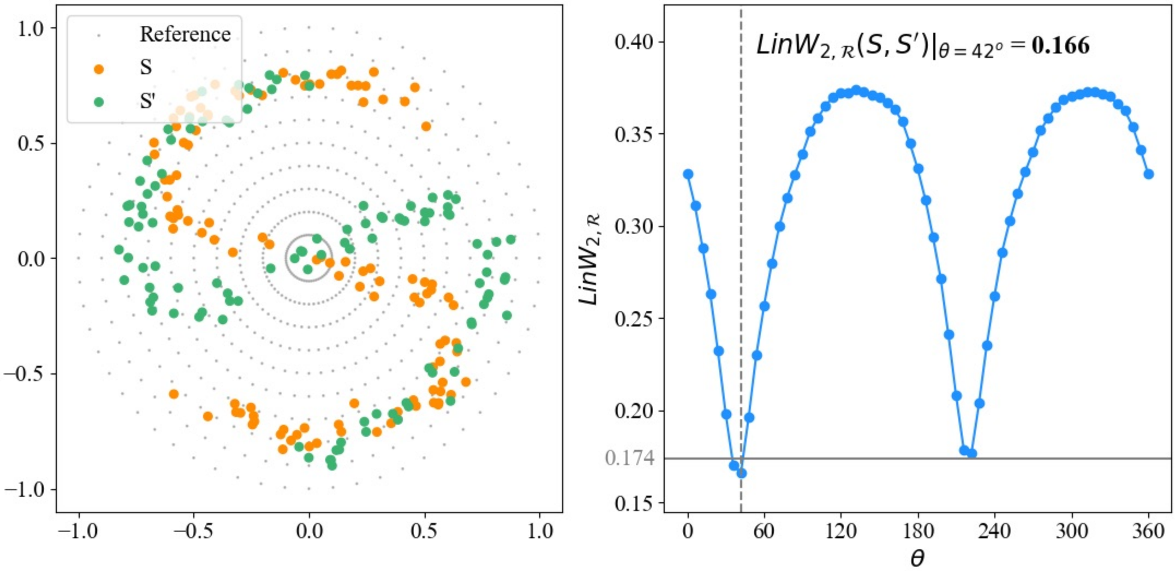


Figure 3.15: *Left*: Two randomly sampled discrete distributions of the letter "S". The green distribution S′ is rotated 43° relative to the upright orange distribution S. The gray reference consists of 10 concentric rings with an angular resolution of 6°. *Right*: The $\mathrm{LinW}_{2,\mathcal{R}}$ pseudo-distance between S and S′ at every 6°. The minimum distance is obtained when $\theta = 42^o$ (gray vertical dashed line), giving an OT value of 0.166 lower than 0.174—the $\mathrm{LinW}_{2,\mathcal{R}}$ pseudo-distance between the two upright S's (gray horizontal line).

angle. More generally, we would like to include other types of symmetries as well such as reflections, in which case the present OT framework may need further upgrades.

Events have multiple scales: OT for Multi-Scaled Distributions

The second complication for event-level analysis is that, unlike a single jet, an event has manifestly multiple scales. In other words, an event consists of several jets (and most likely other objects as well) located at potentially distant parts of the detector, whereas each jet itself may contain hundreds of particles all clustered within a small region. If an optimal transport distance is to be defined directly on an event as a collection of all its constituent particles, then the large separation between jets would wash away any finer substructure within a jet, rendering the overall distance less informative. Similarly, the same loss of information arises if an event is represented only by its jets, i.e., if only jet-level information is used.

What we need here is a more general OT-based distance that can handle multi-scaled distributions, such that both jet substructure and the relative locations of the jets can be incorporated into the distance definition on an (relatively) equal footing. We can accomplish this by modifying the ground metric, or rather the resulting cost matrix between each pair of particles in the ground space. In specific, given two events with n and m jets, respectively, we compute an $n \times m$ *morphology* distance matrix and another $n \times m$ *location* distance matrix, where each entry in both matrices is the corresponding distance between every pair of jets in the two events.

To be more specific, every entry of the morphology distance matrix records the usual OT distance between two jets in each event. Essentially, it looks at the individual jets with their constituent particles after the usual centering and rotating, and then computes the OT distance between all pairs of jets in the two events as before. This tells us how different every jet of one event is from all the jets of the other event, in terms of their morphology, i.e., the internal substructure of the jets.

On the other hand, the location distance matrix discards the particle information and is concerned solely at the jet level. Its entry is defined to be the Euclidean distance between a pair of the jet axes in two events on the collider $y - \phi$ plane. That is, the (i, j)-th entry of the location distance matrix is the Euclidean distance between the i-th jet axis in the first event and the j-th jet axis in second event. Here the issue of event preprocessing strikes back. One would need to make sure that the same event, when rotated with respect to itself, do not give rise to a non-zero location distance matrix.

Now assuming that we have obtained both matrices, the next step is to define a way to combine them into one single distance matrix. For simplicity, our first try is to do an element-wise linear combination of the two matrices, with a tunable weight factor for each. At the end of this step, we would have one single $n \times m$ distance matrix for the two events that incorporates both the morphological information between every pair of jets and the spatial information of how far the jets are located relative to each other.

The final step is to reduce the above matrix to one single number, a scalar that represents the "distance" between the two events. Many methods can be conceived, including using the above distance matrix as the ground cost matrix for an additional optimal transport computation. Of course, in the end, the "best" way to perform the reduction, as well as the method for combining the two matrices in the previous step, are all subject to the judgement of the final performance of the specific task under consideration. Our hope is that these choices would not be too task-specific, and would reveal some generic physics for all collider events.

Events have more than energy flow information: Multi-Species OT

The third point, though perhaps most relevant to full events, also applies to jets. In our current formalism, only kinematics as represented by energy flows are considered in the definition of the optimal transport distance. Charge or flavor information is excluded, even though they can be determined by the experiments. Ideally, one would like to incorporate as many observables as possible. This would require the development of a mathematical framework called *multi-species* optimal transport, where particles of different charges or flavors are treated as different species associated with an additional cost of converting object of one species to another.

As a first try, we are currently focusing on jets in order to avoid the complications of events as mentioned above. The immediate goal is to include particle charge information in addition to energy flows and see whether and by how much it boosts jet tagging performance. Ultimately, we would like to develop a framework that can handle events consisting of a variety of objects, such as a multiple of jets, a few muon hits, some electrons, and etc. This would necessitate crafting a multi-species OT on multi-scaled distributions with invariances, that is, combining all the aforementioned upgrades into a *grand* OT distance for the study of collider events.

Once such a grand distance is available, we can accomplish many tasks with ever increasing power. Potential applications include anomaly detection, detector unfolding, as well as inference of the underlying theory parameter(s) by quantifying how close two theories are in a precise way. This last possible use case leads us naturally to the following chapter, where a different physics scenario is considered for the application of optimal transport. Yet, the underlying statistical problem is essentially the same, allowing one to carry over all the analysis tools developed in one context to another.

Chapter 4

Optimal Transport for Dark Matter Astrophysics

One of the most pressing issues in fundamental particle physics is the identity of some 85% of the matter in the universe. The existence of this dark matter (DM) has long been inferred from its gravitational interactions with normal Standard Model particles. However, its fundamental nature and interactions remain unknown to date. While the LHC continues to hunt for DM—mainly through the inverse process where SM particles collide to produce dark matter, a growing abundance of astrophysical studies have proven critical to the study of dark matter. At this cosmic frontier, DM search is unfolding across a variety of experiments.

Direct detection experiments look for nuclear recoils due to rare interactions of SM particles with DM in the local halo. One example is the LZ Dark Matter experiment [118] deep underground in South Dakota, which primarily hunts for cosmic Weakly Interacting Massive Particles (WIMPs)—the most common hypothesized dark matter particles that make up our galactic halo. At its heart, LZ consists of a large liquid xenon time projection chamber sensitive enough to detect low energy nuclear recoils. LZ and similar direct detection experiments have put stringent constraints on the cross section and mass of WIMPs and other potential dark matter candidates [119, 120].

In contrast, indirect detection experiments search for the annihilation of DM into SM particles by looking into the Milky Way and other galaxy halos. These stable products, presumably from dark matter annihilation or decay, can be tracked by a multiple of "messengers", including neutrinos, charged cosmic rays, gamma rays, X-rays, micro waves, and radio waves. Virtually any astrophysical observation, be it ground-based telescope or balloon-borne detector, can be turned into a dark matter indirect detection search, as they hunt for regions in the sky with excess SM particles. For recent reviews, see [121, 122].

All the above mentioned experimental probes hinge on the assumption that DM also interacts with normal matter through forces besides gravity. However, we currently only have concrete evidence for its gravitational interactions. Another possibility therefore is to directly exploit the DM gravitational interactions at the galaxy scale to infer its properties at the microscopic scale. Such macro-to-micro connection is possible thanks to the fact that small perturbations of the DM distribution in the early universe still manifest themselves in the dark matter halos surrounding today's galaxies. Therefore, a close examination of galactic halos is a peak at this mysterious matter.

To achieve the goal of inferring DM microscopic nature through halo properties, we need to compare astrophysical observations with most intricate numerical simulations of galaxy formation. Such N-body simulations can trace structure formation all the way from the birth of the universe to the formation of halos and galaxies in the current cosmological era. They generate predictions for dark matter density and velocity distributions within galaxies, based on the underlying DM properties specified as inputs. One can then compare these simulations with actual observations to infer which values for the DM properties best match the reality.

Chapter 4 is organized as follows. Section 4.1 motivates the examination of halos as a probe of the microscopic properties of dark matter, especially its self-interaction, and suggests optimal transport as a suitable tool for comparing dark matter halos with rich substructure. In Section 4.2, we rephrase the physics problem in statistical language, introducing the general framework of *simulation-based inference* (SBI), also called *Likelihood-free Inference*. Here we point out an alternative route different from the current mainstream neural-network approach to SBI, laying special emphasis on a particular upgraded version called Bayesian Optimization of Likelihood-free Inference (BOLFI). Details of the BOLFI framework are expounded, where we show how to naturally incorporate optimal transport into the BOLFI framework so as to circumvent the issue of handpicking summary statistics.

Section 4.3 and Section 4.4 give a first demonstration of the proposed statistical framework, focusing on the inference of the disc mass fraction and the halo mass of Milky Way (MW) like halos. Both positive and negative preliminary results are shown and we do not shy away from discussing observed issues. This paves the way for future investigations both of the physics of halo simulation and of the statistical framework of BOLFI+OT. As this project is still ongoing at the moment of writing, the results here represent only our current state of knowledge and major modifications are possible after further research.

The present work, to the best of our knowledge, offers the first systematic application of optimal transport in comparing dark matter halos with rich substructure. We would also like to point out that the novel statistical framework of BOLFI+OT may enjoy much wider applications, both within and beyond the context of dark matter astrophysics. We look forward to further refining the framework and exploring its potential usages elsewhere, including collider physics.

4.1 Probing Dark Matter Nature via Halo Properties

Similar to the Standard Model of particle physics, cosmology has its own standard paradigm consisting of the known SM particles, a cold, collisionless dark matter (CDM) participating only in gravitational interaction, and a cosmological constant Λ associated with dark energy, all encompassed within the theoretical framework of general relativity. This phenomenological ΛCDM paradigm is exceptionally successful at predicting the large scale structure of the universe on distance greater than $\mathcal{O}(\mathrm{Mpc})$, as well as its evolution with time.

In the ΛCDM paradigm, dark matter halo—a clump of dark matter bound by gravity—plays an essential role in the formation of cosmic structures. Seeded by primordial fluctuations, initial matter overdensities collapse under gravitational instability, giving rise to clusters of baryonic and dark matter. As cold dark matter is similar to a pressureless fluid, it collapses faster than baryonic matter and is thus more dominant. Therefore, dark matter halos, as they merge with each other and grow in size, provide the gravitational potential that attract and capture ordinary matter such as gas and dust, leading eventually to the birth of galaxies, galaxy clusters, and other cosmological structures. At the same time, a large dark matter halo may draw many smaller halos (called subhalos) into its own potential well. Those subhalos in turn host their own satellite galaxies, and the entire subsystem orbits around the central halo and its host galaxy. This hierarchical nature of structure formation suggests an incredibly rich substructure of dark matter halos, indicating the possible applicability of optimal transport as a tool to mine such substructure; more to come later.

Typically, a current dark matter halo is confined within at most several Mpc. On such smaller scales below ~ 1 Mpc, the cosmological structure is highly non-linear and

N-body simulations with the standard ΛCDM halos becomes, unfortunately, inconsistent with astrophysical observations. Among the most well-studied challenges to ΛCDM are the core/cusp problem, the missing satellites problem, and the too-big-to-fail problem (TBTF).

The core/cusp problem [123, 124] is concerned about the dark matter density profile in the center of the halo. Whereas ΛCDM predicts a steeply rising central density profile (a "cusp") of $\rho \propto r^{-1}$, astrophysical observations of dwarf galaxies and low surface brightness galaxies, on the contrary, almost universally suggest a constant density core $\rho \propto r^0$. The core/cusp problem persists to date: indeed we are now observing increasingly diverse density profiles, both on galaxy scales and even on cluster scales [125, 126, 127].

The missing satellites problem [128, 129] points out sharply that the number of observed satellite galaxies of the Milky Way is far less than what is forcasted by the ΛCDM model. The paradigm predicts a halo mass function scaling roughly as $dN/dM \propto M^{-1.9}$ [130, 131, 132], all the way down to a minimum mass possibly around the mass of the Earth. This indication of a much bigger population of low-mass halos than large-mass ones is clearly at odds with observation. For example, whereas ΛCDM expects thousands of subhalos that could potentially host galaxies within 300 kpc of the Milky Way, only $\sim$ 50 satellite galaxies are confirmed by observation. Of course, with the advent of increasingly advanced wide-field digital sky surveys, more and more ultra-faint dwarf galaxies have been identified as MW satellites, and it becomes hopeful that the missing satellites problem may even be resolved within the framework of ΛCDM itself.

Finally, the too-big-to-fail problem (TBTF) [133] questions the lack of galaxy formation in the most massive subhalos, despite the normal formation of galaxies in smaller subhalos. In other words, those giant subhalos should be "too big (massive) to fail" to incubate stars and the fact that their luminous counterparts are missing is a puzzle for the ΛCDM model.

The above problems have ignited decades of research looking for remedies to the otherwise highly successful ΛCDM model. One hope is that incorporating realistic baryonic feedbacks into the standard N-body simulations may alleviate some of the aforementioned issues. For example, one would expect supernova bursts to noticably modify the density profile of a halo. However, baryonic processes in astrophysics are not only elusive to precise quantitative understanding; their computational implementation also proves to be highly nontrivial. An entire community is dedicated to the research of such hydrodynamic simulations, which consume exceedingly large computational resources. A single hydrodynamic simulation can takes months to complete! It remains to be seen whether baryonic process alone is enough to account for all the observed discrepancies, if we confine ourselves solely within the ΛCDM paradigm.

A more theoretically tantalizing possibility is that the ΛCDM model no longer holds on smaller galactic scales. In this case, one would need new physics Beyond the Standard Model to provide us with additional particles that can potentially serve as a DM. Fortunately, many theories that have been put forth over the years to tackle the limitations of the Standard Model automatically propose the existence of new particles which can serve as dark matter candidates. A well-known example is the aforementioned weakly-interacting massive particles (WIMPs). Motivated by the hierarchy problem, WIMPs has been under extensive search, though with null results so far.

Generic in many BSM physics models is the existence of a dark sector that parallels the familiar SM dynamics. In such dark sectors, DM particles can scatter with each other through $2 \to 2$ processes. Such self-interacting dark matter (SIDM) was originally proposed more than two decades ago as an attempt to resolve the core/cusp and missing satellites problems [134] and still remains an encouraging alternative to the standard paradigm of a collisionless cold dark matter. By allowing energy and momentum transport, dark matter self interactions give rise to significant deviations from the CDM predictions of the structure and dynamics of halos, especially towards its inner region. At sufficiently large radii though, SIDM halos display the same structure as CDM halos since the collision rate is negligible there with a very low DM density. This convergence is not just desirable but requisite, given the observational success of the ΛCDM model on large scales.

As the resolution of N-body simulations improves, we are now able to probe small scale structures key to the differentiation of various physics models of dark matter, where even the subhalos contained in each host halo can be accurately resolved. We now review several basic properties of dark matter halos and galaxy formation, alongside with the introduction of the specific halo simulator used in our later study.

4.1.1 Halo Properties and Halo Simulation

Once a halo (or a galaxy) stops expanding or collapsing, it has reached an equilibrium point of gravitational stability. The resulting system is said to be virialized. The virial radius R_{vir} of the system, that is, the radius within which the virial theorem applies, is defined as the radius at which the density is equal to $\Delta\rho_{\mathrm{c}}$, where ρ_{c} is the critical density of the universe and Δ is the virial overdensity parameter.

One can relate the virial radius, virial mass, and virial velocity of the system via the following set of equations [135],

$$M_{\mathrm{vir}} = \frac{4\pi}{3} R_{\mathrm{vir}}^3 \Delta\rho_{\mathrm{c}}, \tag{4.1}$$

$$V_{\mathrm{vir}} = \sqrt{\frac{GM_{\mathrm{vir}}}{R_{\mathrm{vir}}}}. \tag{4.2}$$

In a sense, all three parameters $M_{\mathrm{vir}}, R_{\mathrm{vir}}, V_{\mathrm{vir}}$ are equivalent labels of mass. The specification of one determines the other two, once a particular definition for the overdensity parameter Δ is assumed. A common convention is to choose $\Delta = 200$, in which case the corresponding virial radius and virial mass are labeled as R_{200} and M_{200}, respectively. Other choices are equally valid, as long as one remains consistent. In our study, we follow the $\Delta = 200$ convention and set our fiducial mass of a Milky Way-like halo to be $M_{200} = 10^{12} M_{\odot}$ at the present time $z = 0$.

Within the ΛCDM model, the density profile of a dark matter halo is modeled with the Navarro–Frenk–White (NFW) profile [136] via

$$\rho(r) = \frac{4\rho_{-2}}{(r/r_{-2})(1 + r/r_{-2})^2}, \tag{4.3}$$

where two additional parameters r_{-2} and ρ_{-2} are introduced. We define r_{-2} to be the radius where the log-slope of the density profile is 2. This point marks the transition from an inner r^{-1} cusp to an outer profile of a steep fall-off with r^{-3}. Accordingly, ρ_{-2} sets the density at $r = r_{-2}$ [135]. To first approximation, the NFW profile offers a good depiction of the internal matter distribution of a dark matter halo of any mass. Note that baryonic processes are not accounted for here, which can cause the density profile of a halo to deviate significantly from the NFW profile and other upgraded functional forms. It is therefore important that a good halo simulator takes into account such baryonic physics.

Given a halo mass M_{200}, one can replace one of the two parameters in the NFW profile by the halo concentration defined as $c_{200} = R_{200}/r_{-2}$. A combination of M_{200} and c_{200} then completely determines the dark matter density profile of a halo. Concentration displays considerable variation from halo to halo, partially due to the difference in the specific history of halo mass accretion [137]. Qualitatively speaking, early-forming halos, which also tend to have lower masses, assemble at a time when the universe has a higher mean density. Therefore, they usually acquire higher concentrations than larger-mass, later-forming halos, which are hierarchically built up by merging smaller halos according to the ΛCDM narrative. Similarly, many other key halo properties are impacted by the specific astrophysical merger history, making the incorporation of large *halo-to-halo-variance* a critical component of any realistic simulator.

Although dark matter halos can be indirectly observed via methods such as gravitational lensing, the most obvious things we see with telescopes are always their luminous counterparts. Therefore, it is essential to link the properties of observable galaxies with those of the halo that encompasses them. This galaxy-halo connection is extremely complicated. The usual practice is to employ some forward model for galaxy formation from the underlying physics within the standard ΛCDM paradigm or its modified version. We expect a versatile galaxy simulator to incorporate different forward models and provide good empirical relations for the galaxy-halo connection.

As emphasized above, a diverse array of baryonic effects need to be taken into account for a simulator to reasonably model an actual halo. Such baryonic physics is usually “painted” on top of dark matter only N-body simulation. One important and obvious thing to add is a galactic disc at the center of the host halo. Primarily composed of visible matters including stars and gas, a central disc can cause drastic suppression of substructure formation of the host halo. Naturally, the mass of the disc relative to that of the halo is an important property and is usually encoded by the parameter f_m—the disc mass fraction—at the present time $z = 0$. For a MW-like halo, we have $f_m = 0.05$ as the fiducial value.

The halo simulator employed in our study is a semi-analytical satellite galaxy generator called `SatGen` [138, 139], which is designed to be a powerful and fast emulator of much more computationally expensive N-body and hydrodynamic simulations. `SatGen` is able to produce large samples of MW-like halos with high resolution, meaning that it can track subhalos all the way down to a mass of $10^{7.75} M_\odot$. We now briefly describe the process of halo generation using SatGen; for a more complete description, please refer to the original papers [138, 139].

SatGen simulation consists of two major steps in series. First, it generates a merger tree which traces the formation history of a dark matter halo through its accretion and merging processes with other halos [140]. This step is named TreeGen, which is relatively efficient with one merger tree of a $M_{200} \approx 10^{12} M_\odot$ halo produced in about a minute. A number of pre-specified parameters are required to set the resulting properties of the generated halo. For example, M_{200} is obtained from the mass of the main branch (i.e., the most massive progenitor) of the merger tree and can be tuned at this step. The merger tree provides the backdrop against which satellite galaxies are to be evolved in the next step.

Effects from baryonic feedbacks are crucial for the formation and evolution of a large population of satellite galaxies in the host halo. In this second step named SatEvo, the simulator integrates the orbits of each satellite galaxy within a composite potential, accounting for the gravitational interactions between the satellites and the host galaxy. At the same time, tidal forces and other structural evolutions experienced by the satellites are also incorporated. Not surprisingly, such satellite evolution is computationally very demanding, where more than 2 hours are needed to evolve the satellite galaxies for just a single host halo of mass $10^{12} M_\odot$. The required simulation time keeps increasing with larger host halo mass.

Naturally, properties related to baryonic physics such as the galactic disc are inputted in the second step SatEvo. In particular, the disc mass fraction f_m can be tuned here, independent of the previous merger tree step. In practice, this translates to the freedom to scan a range of f_m values given one same halo merger tree.

4.1.2 Optimal Transport on Halo Substructure

The rich substructure of a dark matter halo encodes many of its key properties, such as the mass, the fraction of its central disc, and the particular hierarchical merging history, among others. At the same time, it also reveals the essential dark matter microscopic properties universally governing all halos. The hope is that a principled comparison between different halos (for example, simulation *vs.* observations) can inform us more on the general particle nature of dark matter than on the specific astrophysical history and properties of one halo. In particular, as dark matter self-interactions can drastically modify the phase space distribution of a halo depending on the strength of the interactions, it seems viable and promising to constrain the DM self-interactions via a halo-to-halo comparison, which is the ultimate dream of our project.

But first, one would need a quantitatively precise way to define how similar two halos are based on their substructure. This is where optimal transport comes in—to replace the traditional summary statistics which simply reduce a halo down to a few scalars. Indeed, the present situation is very similar to the collider physics scenario. Here the distributions under comparison are dark matter halos instead of collider jets. If optimal transport can again provide a useful metric structure, then the comparison between dark matter halos will be straightforwardly facilitated by a precise OT distance calculation, where halos with smaller OT distance will (hopefully) share similar underlying properties.

Of course, this optimal transport idea may not work at all in the present case of dark matter halos. Essential to its success is the input distributions themselves. In other words, we need to think of a most informative way to represent a halo as a distribution. In the case of collider physics, the natural way to represent a jet as a p_T-weighted discrete distribution on the $y - \phi$ plane turns out to be extremely effective for tagging and many other applications. This "miracle" does not just happen on its own, but is deeply grounded in theoretical reasons; see [13] for a pioneering exploration into the topic.

On the other hand, it is far from clear what distribution one should choose to represent a dark matter halo. Both the ground space and mass of the distribution need to be properly defined in order to encode the essential physics about the underlying parameter. This also implies that the criteria of a distribution being informative may well differ according to which theory parameter is under consideration: a specifically weighted distribution in a certain ground space that works best for the inference of, say, the halo mass does not necessarily assure the success of inferring the self-interaction strength.

In any case, it is reasonable, or rather only possible, to use features that can be extracted from simulation and observational data as the ground space and mass for the OT formulation. In specific, the velocities and radial distances of the visible satellite galaxies are good choices as a first try, since they can be easily obtained from observations and simulations and should encode halo substructure information to certain degree. We therefore propose to construct three different 2-dimensional ground spaces using the pairs (r, v_{total}), $(r, v_{\text{tangential}})$, and (r, v_{radial}), where r is the distance from each satellite galaxy to the halo center, v_{total} is the magnitude of the total velocity of the satellite orbiting the central halo, and $v_{\text{tangential}}, v_{\text{radial}}$ are the tangential and radial parts of v_{total}, respectively. In practice, we always normalize the two ground space axes so that their ranges are both on the order of $\mathcal{O}(1)$ and therefore contributions from the two features (i.e., r and v's) are roughly equal. [1] This motivates the choice of a HK distance with $\kappa = 1$, and in the following study we exclusively focus on $\text{HK}_{\kappa=1}$ for all the presented results. [2]

As for the "mass" of the individual satellite galaxies of a host halo, we pick for our

[1] For the ground metric, we have tried both the Euclidean distance and the Manhattan distance on the 2D ground space and observed no noticeable difference in the resulting OT distances computed. Therefore, we stick with the Euclidean distance as the ground metric for all following studies.

[2] We have also tried different OT distances with κ values ranging form $+\infty$ (W_2), to $100, 10, 1, 0.1, 0.01$. Our preliminary results suggest HK distance with $\kappa \sim [0.5, 1]$ as the best performing OT distance for the tasks currently considered. Further studies are needed to examine the effect of tuning κ on the downstream statistical task.

current preliminary study two natural choices: the log of the mass of each satellite galaxy, $\mathrm{Log}(M_{\mathrm{sat}})$ [3], and the absolute value of the v-band magnitude of each satellite, i.e., its luminosity L [4]. Therefore in total, there are six different distributions to represent the substructure of one host halo; see Figure 4.1 for an illustration for two random halos both with $M_{200} = 10^{12} M_{\odot}$ and $f_m = 0.05$, where two are chosen in order to demonstrate intrinsic halo-to-halo invariance due to different merger histories.

Now assuming optimal transport with the above ground spaces and mass definitions works as we expect, we can proceed forward to use OT distances for the inference of the underlying theory parameter(s). This problem belongs to the statistical field of Simulation-based Inference (SBI), which is the topic of the next section.

4.2 Bayesian Optimization for Likelihood-free Inference

Suppose we are interested in constraining the mass of the observed Milky Way halo by comparing it to a set of simulated MW-like halos. Here, the halo mass defines our *theory parameter*, denoted as θ. In principle, one can have multiple theory parameters $\boldsymbol{\theta}$ that are to be inferred simultaneously. The method presented below works well when $\boldsymbol{\theta}$ is low dimensional. When the number of theory parameters gets large, additional tricks may be required to maintain the efficiency of the inference framework. For simplicity, let us stick with one theory parameter, which will be the case for our later applications.

Now given a specific value of θ drawn from some prior distribution $p(\theta)$ (usually

[3] We need to take the logarithm of the mass as it may vary approximately from 10^4 to 10^9 for different satellite galaxies. If we were to directly use the galaxy mass, then the OT distance would be completely dominated by the most massive satellites.

[4] The magnitude is a unit for stellar brightness, and the v-band means that the passband is within the visible light range. Since the v-band magnitude is negative for the galaxies in our dataset, we use its absolute value as the "mass".

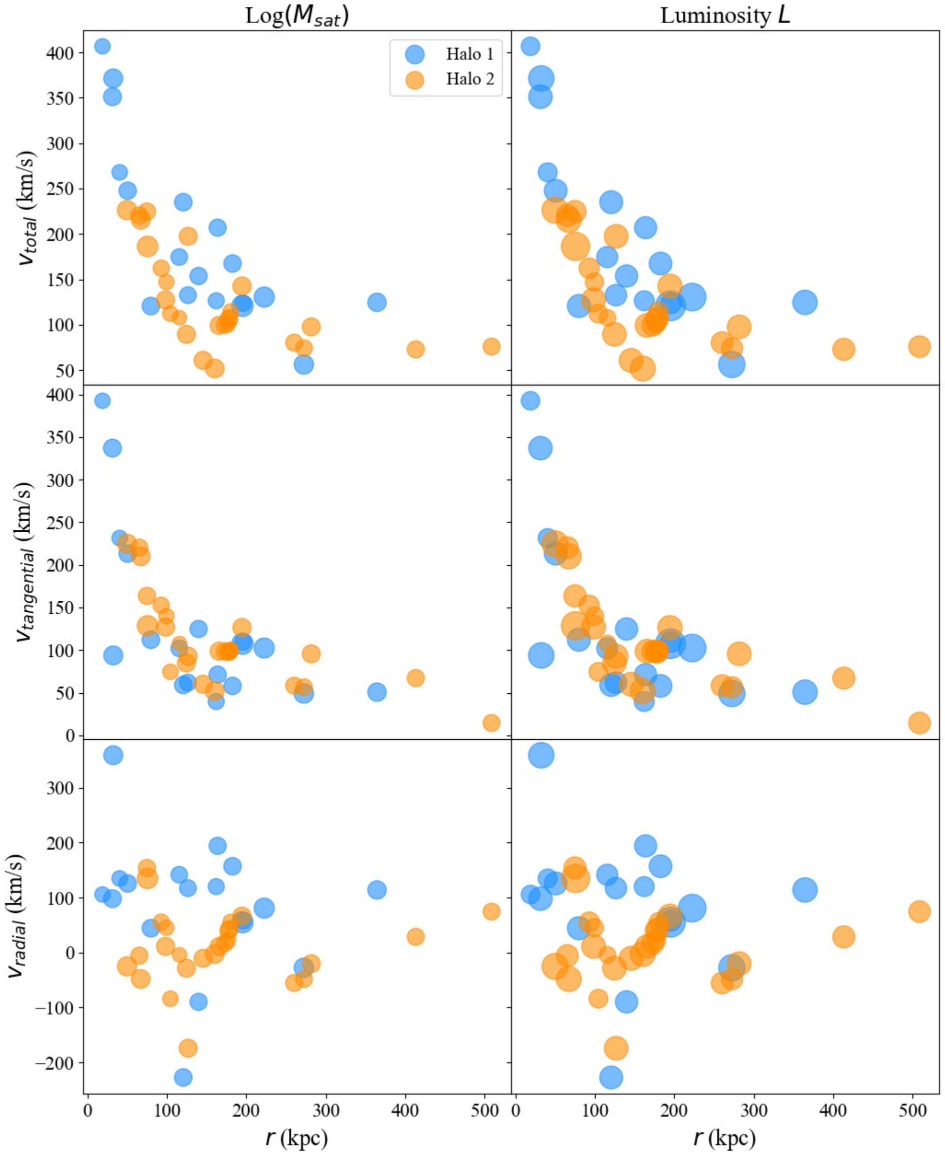


Figure 4.1: Two host halos (one orange, one blue) with $M_{200} = 10^{12} M_\odot$ and $f_m = 0.05$ simulated by `SatGen`, with their respective satellite galaxies displayed in three different ground spaces under two different "mass" choices. *Top to bottom*: the ground space for OT computation is chosen to be (r, v_{total}), $(r, v_{\text{tangential}})$, and (r, v_{radial}), respectively. *Left to right*: the mass for individual satellite galaxy is Log(M_{sat}) and luminosity L.

a uniform distribution), the simulator then generates certain outputs based on some complex physical processes it is designed to model. Of course, the simulator most likely involves many other underlying parameters in addition to θ—the primary parameter of interest. We refer to those parameters as the *latent variables* z. Further, as the simulator oftentimes includes stochastic components, its output observables are typically not in a simple one-to-one correspondence with the input parameters. Therefore, given specific inputs θ and z, we model the simulated outputs x with some probability $p(x|\theta, z)$. In the case of inferring Milky Way mass, the theory parameter θ is the halo mass; one important latent variable z is the galaxy-halo connection; the simulated output x is a specific host halo with certain subhalo distribution; and the actual observation x_o is the, of course, Milky Way itself.

In general, this forward direction from the theory parameter to the simulated observables is straightforward and well grounded in physics. One runs the simulator and the software "automatically" gives the desired outputs. However, it is often the *inverse problem* that is of real interest in science. One usually is not content with just getting some simulated data, but would like to go a step further to compare the simulation with the ground truth in order to constrain the underlying parameter of the actual observation x_o. In other words, our primary goal is to obtain the posterior distribution $p(\theta|x)_{|x=x_o}$, conditional on the observation. Bayes' rule tells us that

$$p(\theta|x)_{|x=x_o} = \frac{L(x|\theta)_{|x=x_o} p(\theta)}{\int \mathrm{d}\theta' \, L(x|\theta')_{|x=x_o} p(\theta')}, \tag{4.4}$$

where $L(x|\theta)_{|x=x_o}$ is the likelihood function of x with respect to θ when we observe $x = x_o$ and $L(x|\theta)$ is defined by

$$L(x|\theta) = \int \mathrm{d}z \, p(x, z|\theta). \tag{4.5}$$

Usually, the high complexity and expensive computational cost of many simulators (including our astrophysical simulators) make it impossible to deduce an explicit, closed-form solution for the above likelihood function. Under such circumstances of an intractable likelihood function, traditional inference methods fail, necessitating the introduction of novel techniques that do not require a known likelihood function. Given the ubiquitousness of these problems in almost every discipline, a tailored field of simulation-based inference is born where diverse methods are used to estimate the posterior without the need to extract the exact likelihood function.

With the rise of neural networks, many recent studies on SBI has turned to NNs, using them to emulate the likelihood function or the posterior distribution directly. Among such NNs acting as surrogates, a prominent one is normalizing flows [141], which evokes a series of invertible and differentiable transformations to turn a complex distribution into a simple one (called based distribution and is usually a Gaussian). Yet just like other NN approaches, flow-based models also suffer from interpretability issue due to its long chains of transformations introducing high dimensional latent spaces. Whenever possible, it is desirable to use easy-to-understand statistical methods in place of NNs.

That is why we turn to a different approach under the general category of Approximate Bayesian Computation (ABC). First described in [142], ABC is nothing new at all and its advantage lies in its conceptual clearness. Without the need to resort to machine learning for explicitly emulating any function, ABC builds upon our intuition that simulations "closer" to the actual observation should share similar theory parameters. Essentially, the vanilla ABC employs a simple rejection sampling scheme. First, we forward simulate a set of data from various values of the theory parameter drawn from the prior $p(\theta)$. Then, if the discrepancy between the simulated data and observed data is smaller than some threshold ε, we keep this simulation; otherwise, we throw it away. At the end of the day, the acceptance rate of the simulations serves as an approximation of the posterior distribution of theory parameter, with the one corresponding to the highest acceptance rate indicating the optimal value of the parameter.

Key to this ABC framework is a notion of discrepancy, which measures how different a given simulation is from the actual observation. In its usual version, this discrepancy score depends crucially on good summary statistics, which usually reduce the data (observation and simulation) to a few low-dimensional features. Expert knowledge (and a bit of luck) is indispensable to a shrewd choice of summary statistics. Indeed, they are exactly the same in spirit as the jet substructure observables discussed in the previous chapter under the collider physics context.

However, when the system under study is sufficiently complex and the data high dimensional, as is the case for our astrophysical application, manually selecting a small set of summary statistics most certainly leads to information loss, as important features may be overlooked. Indeed, it becomes ever more challenging to craft the summary statistics, when the underlying mechanisms governing the system are themselves not fully understood. As the readers may already anticipate, optimal transport provides a promising alternative to the usual practice of handpicking an *ad-hoc* set of summary statistics. It offers a novel and more sophisticated notion of discrepancy between simulation and observation, by better capturing the subtle patterns, dependencies, and higher-order statistical properties in data that my elude even an expert's eye. Of course, the success of optimal transport still hinge on a proper choice of the ground space with its ground metric and that choice is highly problem specific.

Now back to the ABC framework. Another problem of its vanilla implementation is that the model most often ends up throwing away too many simulations due to the simple uniform cut on discrepancy. This is largely caused by a lack of knowledge about how the theory parameters influence the discrepancy score, which makes the rejection scheme highly inefficient. An exceedingly large number of simulations are therefore required, necessitating a formidable amount of computational time. This issue can be partially resolved by an upgraded version of ABC, termed Bayesian Optimization for Likelihood-free Inference (BOLFI) [143], where a technique called active learning is introduced to vastly improve sampling efficiency for expensive simulators. The main innovation of BOLFI is a combination of an optimization strategy and a probabilistic modeling of discrepancy score with respect to the theory parameters.

In the following subsections, we successively discuss the individual components of the BOLFI framework, where examples and plots are give to illuminate the statistical formulation. Figure 4.2 presents a schematic flowchart of the BOLFI+OT framework to guide our conceptual understanding.

4.2.1 Gaussian Process Regression

To reiterate, given an observation x_o, our goal is to infer the theory parameter $\theta_o \in \mathbb{R}$ that gives rise to this observation. According to the ABC recipe, we sample a set of theory parameters $\{\theta_1, \theta_2, \cdots, \theta_k\}$ and generate the simulated data $\{x_1, x_2, \cdots, x_k\}$ corresponding respectively to each θ_i. We then define the discrepancy score between observation x_o and simulation x_i, either by choosing some summary statistics or by evoking optimal transport distance. We denote the resulting discrepancy score as $\{\Delta(\theta_1), \cdots, \Delta(\theta_k)\}$, where the parenthesized θ_i's highlight the dependence of discrepancy on the theory parameter. Intuitively, as the discrepancy score Δ decreases, we expect the simulation to become increasingly similar to the observation. In other words, the parameter $\hat{\theta}$ that minimizes discrepancy will likely be close to θ_o that we wish to infer. Eventually, everything boils down to finding a relationship between the discrepancy score Δ and the theory parameter θ, whose exact functional form may be impossible to write down.

Now BOLFI proposes to use probabilistic modeling to infer the above relationship.

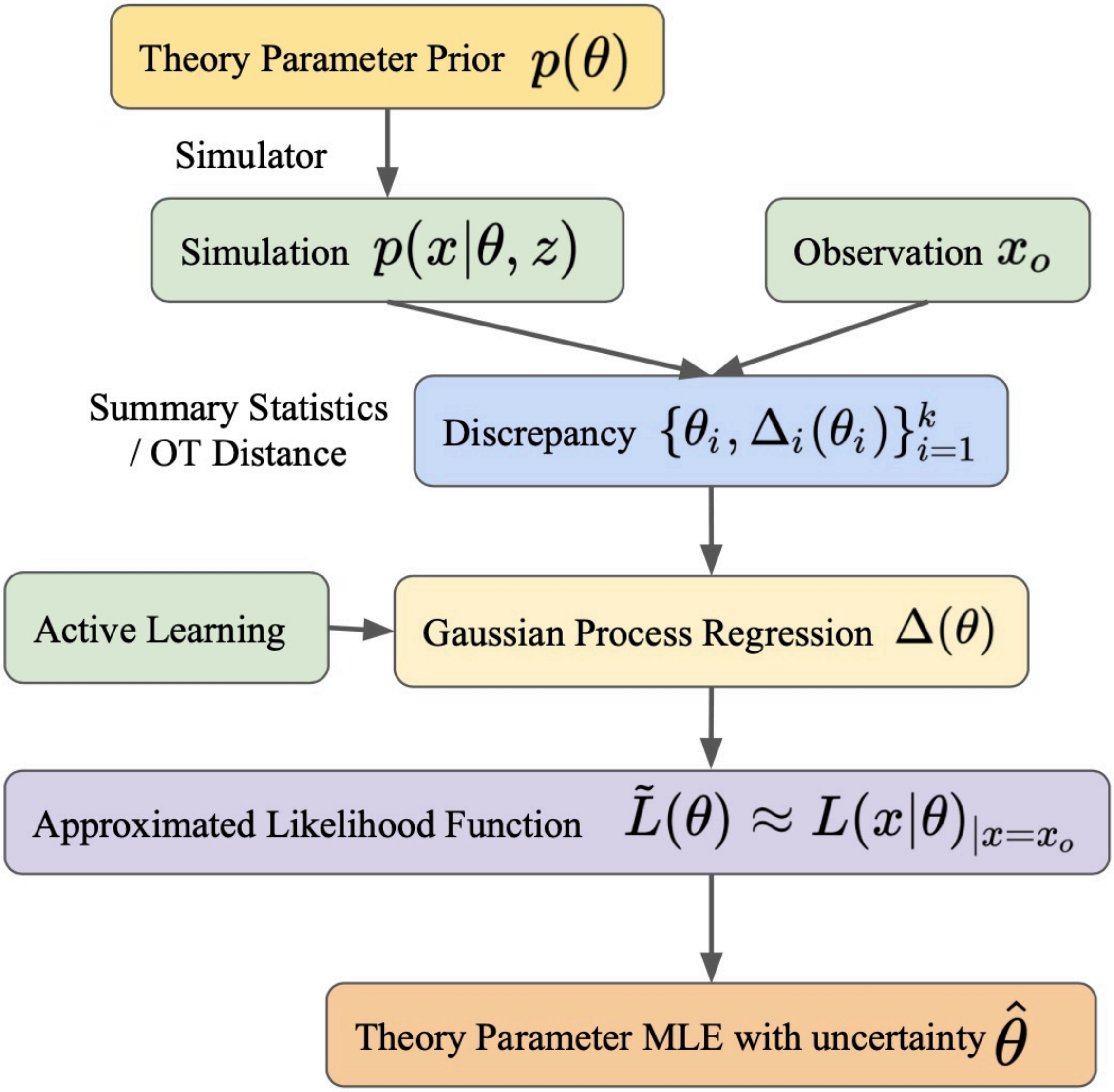

Theory Parameter Prior $p(\theta)$
Simulator
Simulation $p(x|\theta, z)$
Observation x_o
Summary Statistics / OT Distance
Discrepancy $\{\theta_i, \Delta_i(\theta_i)\}_{i=1}^k$
Active Learning
Gaussian Process Regression $\Delta(\theta)$
Approximated Likelihood Function $\tilde{L}(\theta) \approx L(x|\theta)_{|x=x_o}$
Theory Parameter MLE with uncertainty $\hat{\theta}$

Figure 4.2: A schematic flowchart of the BOLFI framework for simulation-based inference problems. The key steps will be explained in the following main text.

This is in essence a regression problem, where the tuples $\{\theta_i, \Delta_i(\theta_i)\}_{i=1}^k$ provide the required set of k training data. A myriad of methods exist for such regression problems, ranging from simple polynomial regression to neural networks. However, the key complication here lies in the fact that due to the stochastic nature of the simulator, one same value of θ may give rise to a variety of simulated data x, which in turns results in different values of Δ. This is why probabilistic modeling is required to take into account this intrinsic data variance and uncertainty.

In practice, BOLFI suggests using a Gaussian process (GP) to model the objective function $\Delta(\theta)$. GP assumes that the set of discrepancy scores $\{\Delta(\theta_1), \cdots, \Delta(\theta_k)\}$ is drawn randomly from a Gaussian distribution, with mean function $m(\theta)$ and covariance function $\Sigma(\theta, \theta')$. It then outputs the value of discrepancy $\Delta(\theta_{k+1})$ as a probability distribution (again a Gaussian) at any new parameter point θ_{k+1}. In a sense, GP offers a recursive relationship for Δ in terms of θ and assigns a certain probability to each possible regression functions $\Delta(\theta)$, where the mean function $m(\theta)$ gives the most probable functional relationship between Δ and θ. Figure 4.3 illustrates how a Gaussian Process fits an underlying function given a training set of input data.

More precisely, using the shorthand notation $\Delta(\theta_{1:k}) := \{\Delta(\theta_1), \Delta(\theta_2), \cdots, \Delta(\theta_k)\}$ and $\theta_{1:k} := \{\theta_1, \theta_2, \cdots, \theta_k\}$, the distribution of $\Delta(\theta_{1:k})$ can be written as

$$\Delta(\theta_{1:k}) \sim \mathcal{N}(m(\theta_{1:k}), \Sigma(\theta_{1:k}, \theta_{1:k})), \tag{4.6}$$

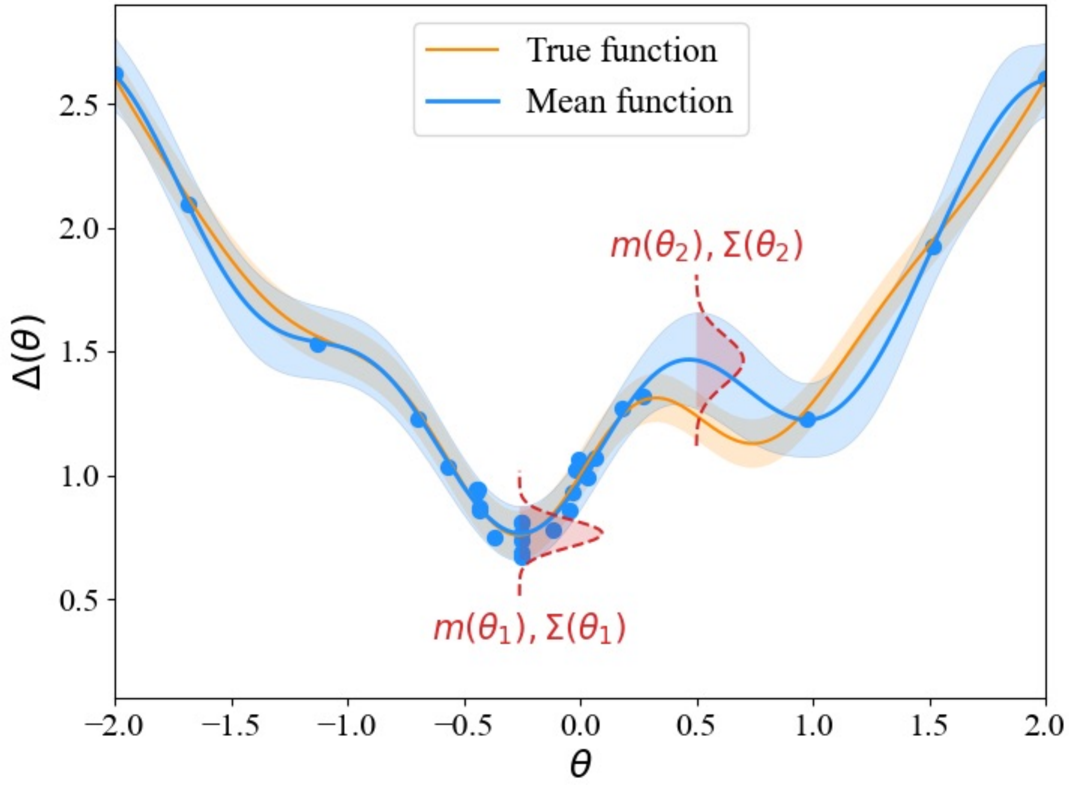


Figure 4.3: An illustration of a Gaussian Process. The orange line represents the true function $\Delta(\theta)$ with a 2σ error band that the GP intends to model. The blue dots are the training data inputted to the GP. The blue curve (with a 2σ blue band) represents the mean function $m(\theta)$ the GP outputs based on the training data. The vertical Gaussian distribution (dashed red) at a given θ value demonstrates the probability the GP assigns to each possible function value Δ at the specific θ.

where $\mathcal{N}$ is a normal distribution with mean

$$m(\theta_{1:k}) = \begin{pmatrix} m(\theta_1) \\ m(\theta_2) \\ \dots \\ m(\theta_k) \end{pmatrix}, \tag{4.7}$$

and covariance

$$\Sigma(\theta_{1:k}, \theta_{1:k}) = \begin{pmatrix} \Sigma(\theta_1, \theta_1) & \dots & \Sigma(\theta_1, \theta_k) \\ \vdots & & \vdots \\ \Sigma(\theta_k, \theta_1) & \dots & \Sigma(\theta_k, \theta_k) \end{pmatrix} + I_k \sigma_n^2. \tag{4.8}$$

Here I_k is the k-by-k identity matrix, and the additional σ_n^2 accounts for the noise in the observed data.

The standard choice for the mean function is a convex quadratic polynomials, and for the covariance function, a squared exponential. That is,

$$m(\theta) = a\theta^2 + b\theta + c, \tag{4.9}$$

with non-negative a, and

$$\Sigma(\theta, \theta') = \sigma_f^2 \exp\left(\frac{1}{\lambda^2}(\theta - \theta')^2\right). \tag{4.10}$$

Here σ_f^2 is the signal variance and λ is the characteristic length scale. In the case where $\boldsymbol{\theta}$ is more than one-dimensional, one simply add a summation to account for each component of $\boldsymbol{\theta}$.

The above GP model has six hyper-parameters, a, b, c, σ_f^2, λ, σ_n^2, to be tuned by maximizing the likelihood of the data. For simplicity, in our later application, we choose the mean function to be constant, i.e., $a, b = 0$, and thus having $m(\theta) = c$.

With k pairs of training dataset $\{(\theta_1, \Delta(\theta_1)), \ldots, (\theta_k, \Delta(\theta_k))\}$, we can then obtain the posterior distribution of $\Delta(\theta)$ at any point θ using Baye's rule [144],

$$\Delta(\theta)|\Delta(\theta_{1:k}) \sim \mathcal{N}(\mu_k(\theta), v_k(\theta) + \sigma_n^2), \tag{4.11}$$

where

$$\mu_k(\theta) = \Sigma(\theta, \theta_{1:k})\Sigma(\theta_{1:k}, \theta_{1:k})^{-1}(\Delta(\theta_{1:k}) - m(\theta_{1:k})) + m(\theta_{1:k})$$

$$v_k(\theta) = \Sigma(\theta, \theta) - \Sigma(\theta, \theta_{1:k})\Sigma(\theta_{1:k}, \theta_{1:k})^{-1}\Sigma(\theta_{1:k}, \theta)$$

When simulated data enjoy little intrinsic dispersion or noise, even a relatively small number of training data can already empower GP to output a satisfactory regression function with low variance, providing a quick estimate of the probabilistic dependence $\Delta(\theta)$. In such ideal cases, one can directly use the GP regression relation to approximate the likelihood function $L(x|\theta)$; skip the following subsection on active learning and jump directly to Section 4.2.3 for the next step in the BOLFI framework.

However, we are usually not so lucky, certainly not in the case of the astrophysical dataset we are interested in. For one thing, due to different merger histories, halos usually display significant variance even if they have the same underlying mass. Therefore, one would need a large number of simulations at every single mass value (and its vicinity) in order to faithfully cover such halo-to-halo variance. It soon becomes computationally formidable to run the simulator at a large number of θ values. A simple strategy to densely populate the parameter space is unrealistic; one needs additional tricks to do better than a random draw of θ from its prior distribution. This is where active learning comes in.

4.2.2 Active Learning

Active learning answers the question of where to simulate next so as to minimize the total number of simulator calls while at the same time achieve the best possible regression function. As explained above, the mean function and covariance of the GP regression can be updated according to Equation (4.11) with the introduction of a new parameter point θ_{k+1}. Since we are only interested in the region around $\hat{\theta}$ that gives the minimum Δ, there is no reason to waste simulation runs for those θ values that are known to result in large discrepancy. We therefore want more sample points near the current minimum in order to improve resolution in this critical region.

On the other hand, there is no guarantee that the current minimum is the true global minimum. This suggests that we should also explore the parameter space where the variance $v(\theta)$ is large. This is the region of low confidence, most likely due to the lack of sampled points there. Therefore, to balanced the two needs—the first of exploitation and the second of exploration, we need to carefully draft an *acquisition function*, the minimization of which would give the next parameter point of inquiry. There is currently an increasing interest in designing various acquisition functions to emphasize different parts of the parameter space. In our study, we use a simple acquisition function known as *lower confidence bound selection criterion* [145] defined by

$$\mathcal{A}(\theta) = \mu_k(\theta) - \sqrt{\eta_k^2 v_k}, \qquad \eta_k^2 = 2\log\left[k^{d/2+2}\pi^2/(3\epsilon)\right], \tag{4.12}$$

where d is the dimension of the theory parameter space, which is just $d = 1$ for our case. Additionally, ϵ is a small constant set to be 0.1. As before, μ_k and v_k are the mean function and the variance obtained from the existing k data points, as in Equation (4.11).

The above acquisition function consists of two parts, the mean μ_k minus a term proportional to the variance v_k. Minimizing the first term μ_k lets us focus on the parameter region where the discrepancy score Δ is the smallest according to the existing k data points. This gives the exploitative “zoom-in” to the current optimal value of the theory parameter. The second term, on the other hand, explores the region where the current variance v_k is large. There, substantial uncertainty remains about what the true discrepancy is, and hence more data samples are needed to reduce our lack of knowledge.

The next point for inquiry, θ_{k+1}, is then obtained by minimizing the acquisition function Equation (4.12). Of course, one can always sample a number of values around θ_{k+1} in one batch instead of one θ value at a time to save some efforts. Furthermore, a stochastic component is usually added to the acquisition function in order to increase the chance of exploring more values of the theory parameter. More precisely, instead of sampling θ_{k+1} at θ_m, the exact minimum of the acquisition function $\mathcal{A}$, we sample θ_{k+1} from a Gaussian distribution that is centered at θ_m with some characteristic length scale λ as its variance. This is implemented in our later applications.

Figure 4.4 shows how active learning and GP regression works together step by step in an attempt to find the minimum of the underlying function $\Delta(\theta)$ same as in Figure 4.3. First, one has an initial training dataset which the GP regression fits. From this, one calculates the acquisition function, whose minimum outputs the next parameter point of inquiry represented by a star. We add this new parameter point and the corresponding discrepancy Δ to the initial dataset, and then run the GP regression again on the augmented dataset. This will give rise to a new acquisition function for the next step, and the whole process repeats itself until some stopping criterion is met.

In Figure 4.4, we see nicely that the GP regression is successively converging to the true function it intends to model. Furthermore, we can also visually confirm that the acquisition function does generate parameter points that are either around the minimum or in the outer region where samples are insufficient. In practice, for the current astrophysical application, we generate around $\mathcal{O}(50)$ halos for any given value of θ (host mass) at each step, in order to better capture the halo-to-halo variance intrinsic to the data.

4.2.3 Approximated Likelihood Function

Regardless of the use of active learning, the GP regression above always yields a mean function $\mu(\theta)$ with variance $v(\theta) + \sigma_n^2$ for the relationship between Δ and θ. Once we have $\Delta(\theta)$ at hand, we can construct an approximated likelihood function $\tilde{L}(\theta)$ for the minimum of $\Delta(\theta)$ via a non-parametric approach. Using a kernel density estimation [146, 147], an approximation of the likelihood function $L(\theta)$ is given by

$$\tilde{L}(\theta) = E[\kappa(\Delta(\theta))], \tag{4.13}$$

where the kernel κ is chosen such that it has a maximum at zero (though the maximum may not be unique). Assuming the discrepancy is always larger than 0, we can use the uniform kernel for convenience, i.e.,

$$\kappa_u(\Delta(\theta)) = \begin{cases} 1 & \text{if } 0 < \Delta(\theta) < h \\ 0 & \text{if } \Delta(\theta) > h \end{cases} \tag{4.14}$$

where the h is an arbitrary threshold. With this uniform kernel, we have

$$\tilde{L}(\theta) \propto P(\Delta(\theta) < h). \tag{4.15}$$

The problem has now been converted to estimating the probability of the discrepancy Δ dropping below the threshold h. The threshold h is a critical choice. In our work, we

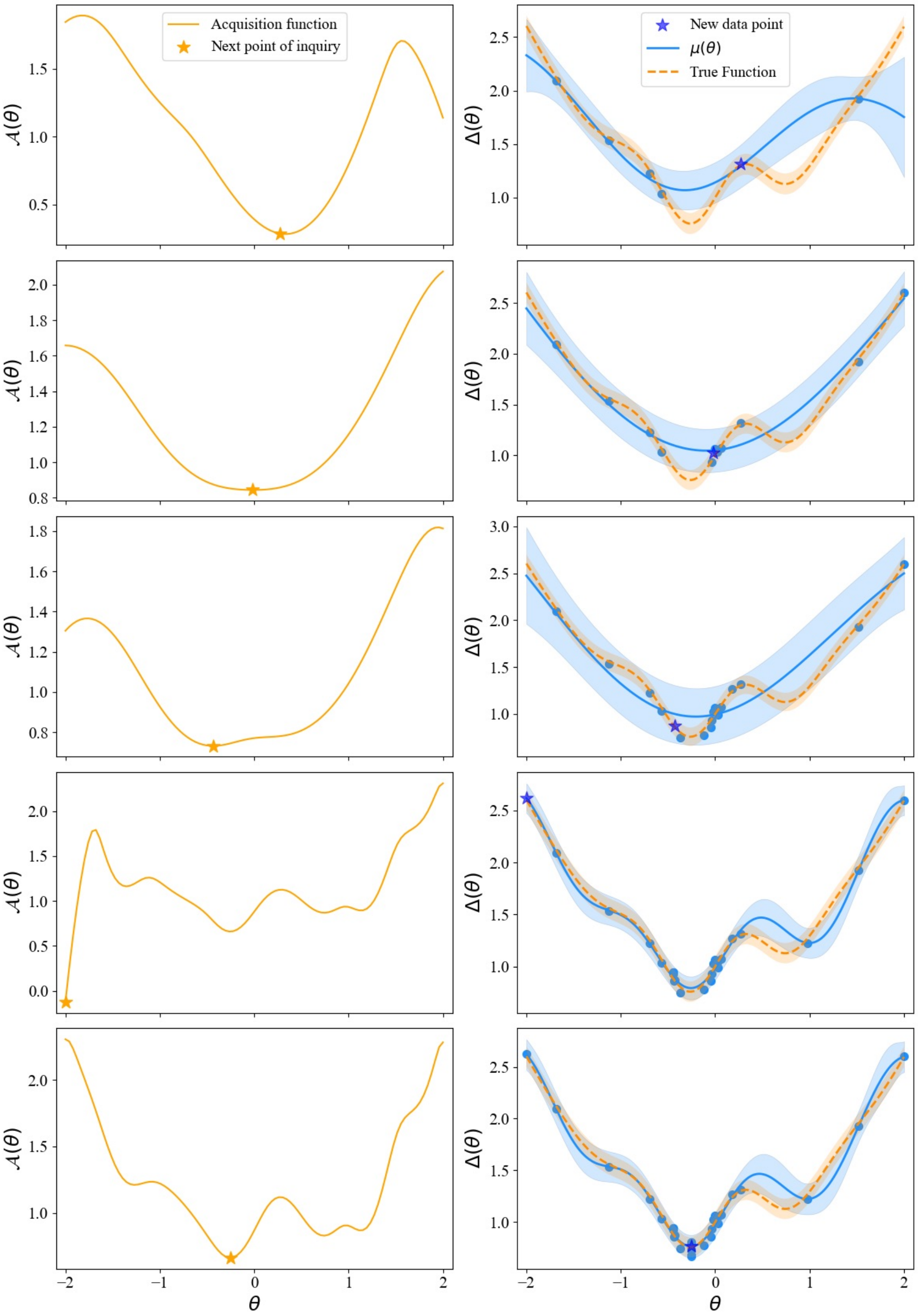


Figure 4.4: Acquisition function (left) and Gaussian Process regression (right) at each active learning step. The dashed orange line (right) represent the true function (with a 2σ band) $\Delta(\theta)$ that the GP intends to fit. At each step, the blue dots are the sample points in the current dataset; the blue line (with a 2σ band) gives the GP regression output; and the starred point highlights the new parameter point to run the simulator determined by minimizing the acquisition function (left, orange).

define h to be the minimum of mean $\mu(\theta)$.

For discrepancy $\Delta(\theta)$ with mean μ and variance $\sqrt{v+\sigma_n^2}$, the likelihood function is given by

$$\tilde{L}(\theta) \propto \mathrm{CDF}\left(\frac{h-\mu(\theta)}{\sqrt{v(\theta)+\sigma_n^2}}\right), \tag{4.16}$$

where CDF stands for the cumulative distribution function of a Gaussian distribution with a mean of 0 and a variance of 1. The approximated likelihood $\tilde{L}(\theta)$ gives a numerical value that represents how well the model, when described by a certain value of the parameter θ, matches the observed data. One would therefore like to maximize $\tilde{L}$ and the corresponding $\hat{\theta}$ then defines the maximum likelihood estimation (MLE) of the parameter. Of course, we also need to assess the accuracy of the estimated MLE. Since the integration of likelihood function by itself does not have statistical significance, we use instead the likelihood region to represent the confidence interval, which is the region of parameter θ that corresponds to the likelihood larger than a certain percentage of its maximum [148]. For example, a 14.7% likelihood region produces a 95% confidence interval.

Figure 4.5 shows the final approximated likelihood function and the corresponding MLE of the θ parameter, obtained from the last step of the GP regression in Figure 4.4 (bottom row, right plot). Again, the true underlying function is the same as before, with its true minimum highlighted for a comparison with the MLE estimated using BOLFI. It is satisfying to see that the MLE is very close to the true minimum [5] and the corresponding 95% confidence interval is narrowly confined, indicating that the model is rather confident in its estimation of $\hat{\theta}$.

Note that it does not really matter whether the GP regression function closely matches the true function over the entire range of θ values, since we are only concerned about

[5] The true minimum is at $\theta = 0.263$, whereas the MLE is $\hat{\theta} = 0.256$ with the lower bound of the 95% confidence interval at 0.174 and the upper bound at 0.338.

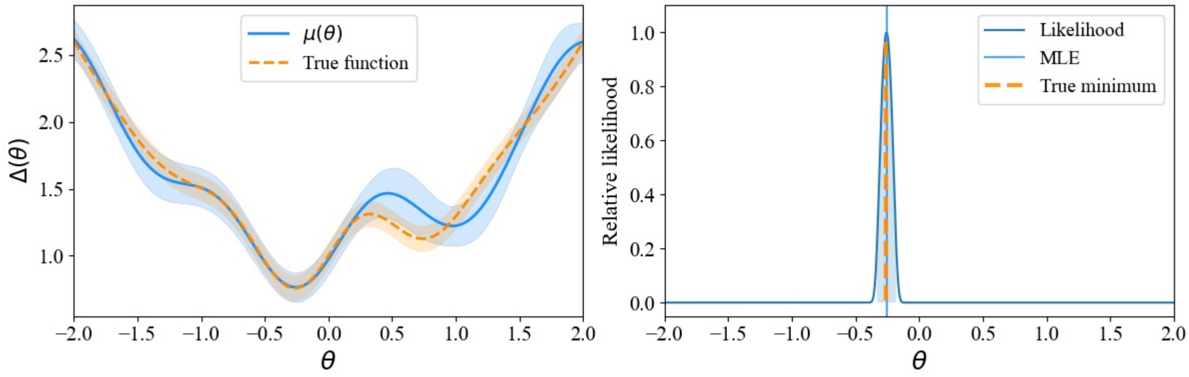


Figure 4.5: *Left*: The final GP regression function $\Delta(\theta)$ (blue) based on all existing data samples, compared to the true underlying function (orange). *Right*: The approximated likelihood function (blue) obtained from the GP regression on the left, and the corresponding maximum likelihood estimation (MLE) of the parameter θ (blue) compared to the true minimum of θ (orange).

the parameter region where Δ obtains its minimum. For example, in Figure 4.5, the GP regression function is rather different than the true function for $\theta \gtrsim 0.5$ due to lack of data samples there (see the bottom right plot in Figure 4.4). Yet, it in no way impacts the excellent inference of the θ value corresponding to the smallest Δ.

The above three subsections conclude the BOLFI framework. As a first try, we now apply the statistical tool to the inference of the disc mass fraction f_m and the halo mass M_{200}, presented respectively in the following two sections. In our current study, only simulated data are used and we generate *mock* observations in the place of real observational data. This way, the ground truths for the underlying parameters are known *a priori*, enabling us to validate our analysis pipeline before we move on to infer the properties of actually observed halos. The hope is that soon our BOLFI+OT framework would be able to help determine halo properties which are poorly constrained at the moment using traditional statistical methods, and then ultimately to have a say on dark matter self-interactions and its general particle nature.

4.3 A Failed Example: Inference of the Disc Mass Fraction of Milky Way-Like Halos

We first set the disc mass fraction f_m as our theory parameter to be inferred. Even without going through all the steps of BOLFI, we will soon discover the failure of this inference task from the regression plots $\Delta_{\rm OT}(f_m)$; see Figure 4.6. Still, we present the study here in order to draw a sharp contrast with the semi-successful inference of M_{200} in the next section.

The disc mass fraction f_m, being related to baryonic physics, is an input parameter to the second step `SatEvo` in the simulator `SatGen`. This means that one can easily tune f_m to any reasonable value and simulate the corresponding halos straightforwardly. In practice, we generate 500 merger trees in the first step `TreeGen`, with the halo mass fixed at $M_{200} = 10^{12} M_\odot$. [6] For each single merger tree, we then put in the satellite galaxies and let them evolve in the second `SatEvo` step for 50 times, each corresponding to one specific f_m value randomly picked from a uniform distribution between 0 and 0.1. [7] In other words, we generate 10 halos at a particular f_m value with each halo corresponding to one merger tree in `TreeGen`, resulting in a total of 500 halos. [8] This small trial dataset is used as a first look at the GP regression of $\Delta_{\rm OT}(f_m)$ in order to determine if the inference task is even possible of success.

[6] Another two important parameters in `TreeGen` are the halo response and the infall orbital parameter distribution, where we set the former to be consistent with NIHAO simulations [149] and the latter to be Zhao-Zhou Li's distribution [150].

[7] When evolving the satellite galaxies, the flattening (disk scale radius / disk scale height) is set to 12.5, the bulge mass fraction is set to 0, and the stripping efficiency of the tide effects is set to 0.6; see [151] for more details.

[8] To account for the fact that observations cannot distinguish very faint satellites, we apply a cut on the surface brightness of all satellite galaxies. The surface brightness is defined as $\mu_v = M_v + 36.57 + 2.5 \mathrm{Log}(2\pi r^2)$, where M_v is the v-band magnitude (a negative value) and r is the half-light radius in kpc. The smaller the values of the surface brightness, the brighter the galaxy is. We require the surface brightness to be smaller than 28 for a galaxy to be detectable. All satellites with a larger μ_v value are discarded from the dataset.

As mentioned above, instead of using actual observational data, for our current study we generate a set of *mock* observations, i.e., simulated halos with the same underlying theory parameter. In specific, for the task of inferring f_m, we set the halo mass to be $M_{200} = 10^{12} M_\odot$ and the disc mass fraction at $f_m = 0.05$ for the 10 halos serving as the mock observations. The reason why multiple halos need to considered rather than a single one is that even with the same theory parameters, halos display considerable variance due to their distinct merger tree histories, which in turn have a large effect on the distribution of satellites within the host halo. Further study must be conducted to determine how many halos are needed to account for this inherent halo-to-halo variance.

The generation of a set of mock observations is essential, since it serves as the reference against which the OT distance is computed. Specifically, to obtain the discrepancy score, we first compute the exact HK distance with $\kappa = 1$ between the individual halos in the dataset and each of the ten halos acting as the mock observations, and then calculate the average of these ten HK distances, which becomes the final discrepancy score Δ_{OT} for each halo with its specific f_m value. Of course, a simple average may not be the optimal way to combine the OT distances when the goal is to incorporate the halo-to-halo variance in the resulting discrepancy score. Other methods will be examined in details in the future.

We now apply the Gaussian Process regression on the discrepancy score as a function of the theory parameter f_m. The OT distance is evaluated on the aforementioned six different pairs of ground space and mass choices. Again, all ground spaces are rescaled, i.e., we divide the values of r and v_{total} (or $v_{\mathrm{tangential}}$, or v_{radial}) for all halos by 100, so as to ensure that the x and y axes of the ground space are of the same order of magnitude ($\sim \mathcal{O}(1)$). This rescaling has no impact on the final inferred MLE. Figure 4.6 shows the resulting GP regressions for the six available ground space and mass choices. As can be seen, all six plots have similar flat distributions for the discrepancy *vs.* disc mass fraction regression, which immediately indicates the failure of this particular inference task: there is no way to infer a minimum value of Δ_{OT} and the corresponding f_m. Therefore no further inference step is necessary and the whole inference task should be aborted.

We can speculate about a variety of possible reasons for this failure. One easily-solvable issue may be simply due to insufficient data, both for the trial dataset and for the mock observations. Improving data statistics, however, is unlikely to resolve the problem, as the current regression relation of $\Delta_{\mathrm{OT}}(f_m)$ seems too flat to indicate any promise. Another possibility is that other OT distances should be considered as the discrepancy score. We have tested a few other κ values and yet observed no noticeable difference. A more physical reason may be that the current six pairs of ground space and mass do not capture the necessary information about the disc mass fraction. One would then need to search for other observable features to use as the ground space and mass.

On a more fundamental level, it may turn out that the dynamics of the satellite galaxies of a host halo (at least as encoded by the ground space and mass choice) is indeed not directly related to the disc mass fraction f_m, although the presence of a massive disc has been shown to enhance the destruction of halo substructure [152, 153]. Still, the large halo-to-halo variance may wash away such difference [139], giving rise to the scattered nature of the data points at each f_m value observed in Figure 4.6. As such variance is intrinsic to the data, it would be extremely difficult to eliminate or even to average out this latent variable, which constitutes one major challenge in astrophysical simulations.

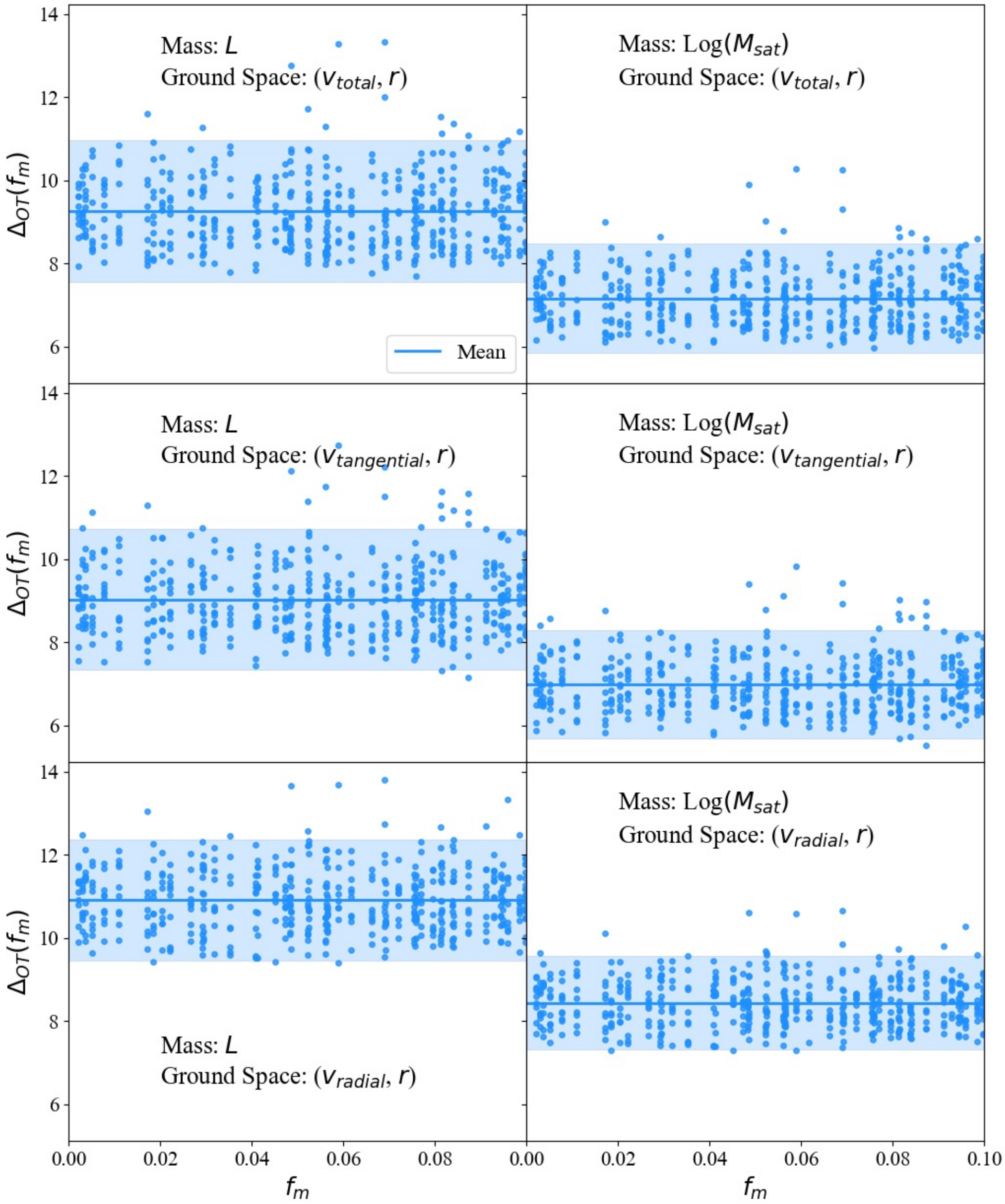


Figure 4.6: Gaussian Process Regression of the discrepancy score Δ_{OT} as a function of the disc mass fraction f_m for 500 simulated MW-like halos with $M_{200} = 10^{12}M_\odot$ and $f_m \in [0, 0.1]$ uniformly. Here Δ_{OT} is obtained by averaging ten HK distances with $\kappa = 1$, each with respect to one halo in the mock observations. Six choices of difference pairs of ground space and mass for the OT computation are studied (indicated by the text in each subplot), where similar regression functions are obtained. The ground truth value for f_m is 0.05. The flatness of the resulting mean function and the scattered randomness of the data points suggest the failure to constrain f_m with the current dataset and method.

4.4 A Semi-Successful Example: Inference of the Halo Mass

We now proceed to infer the halo mass M_{200} using the same method as above. Again, we first generate a smaller trial dataset of 500 halos to facilitate faster inference and to determine if a full BOLFI analysis is necessary. If the inference result using the trial dataset is promising, we then use this dataset as the initial jumping-off point and ask active learning to iteratively acquire new data points, eventually arriving at 2500 halos for the final round of regression and inference.

The 500 halos in the trial dataset have the log of their mass uniformly drawn from the interval $\text{Log}(M_{200}) \in [11.5, 12.5]$, whereas their disc mass fraction f_m is fixed at the fiducial value of 0.05. All other parameters are the same as in Section 4.3. We again have 10 halos serving as mock observations, all with $\text{Log}(M_{200}) = 10^{12}$ and $f_m = 0.05$. The discrepancy score between each halo and the ten mock observations is calculated as before, with the $\text{HK}_{\kappa=1}$ distance playing the role of traditional summary statistics. The OT distance is again evaluated on six ground space and mass choices, with the respective GP regression plots of $\Delta_{\text{OT}}(M_{200})$ showing in Figure 4.7.

Several observations can be made for Figure 4.7. First and foremost, the GP regression functions all show a promising valley shape, in stark contrast with Figure 4.6 where the regressed functions are flat. Here, a clear minimum can be drawn from the valley shape, indicating the potential success of this inference task. Second, all six ground spaces give more or less similar results, with almost no difference between using luminosity L or log of satellite mass $\text{Log}(M_{\text{sat}})$ as "mass". The inference performance is slightly worse if (v_{radial}, r) is used as the ground space, whereas the differences between (v_{total}, r) and $(v_{\text{tangential}}, r)$ are negligible. We therefore do not need to go through the trouble of studying all six ground space and mass choices; instead we focus exclusively on the

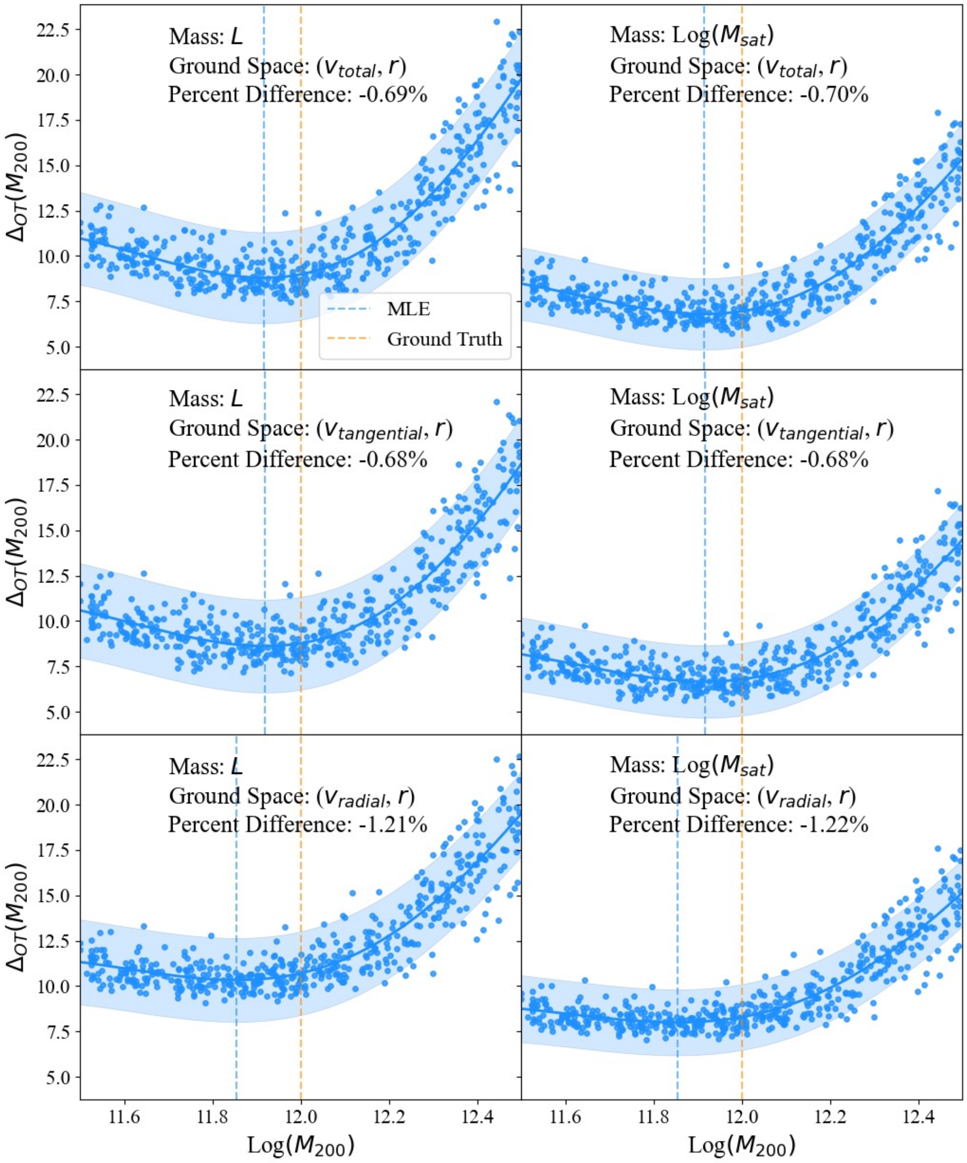


Figure 4.7: Gaussian Process Regression of the discrepancy score Δ_{OT} as a function of the log of halo mass Log(M_{200}), where the unit for M_{200} is $M_\odot$. The trial dataset (blue dots) consists of 500 simulated MW-like halos with $f_m = 0.05$ and Log$(M_{200}) \in [11.5, 12.5]$ uniformly. Here Δ_{OT} is defined to be the average of the HK$_{\kappa=1}$ distances with ten mock observations. Six choices of different pairs of ground space and mass for the OT computation are respectively presented in the subplots. The orange vertical line indicates the ground truth of M_{200} (i.e., $10^{12}M_\odot$), whereas the blue vertical line suggests the Log(M_{200}) value that corresponds to the minimum of the regressed Δ_{OT}. The percent difference between the two are also shown on the plot, with the sign indicating whether the inferred M_{200} is smaller or larger than the true M_{200}.

pair $(v_{\text{total}}, r) + L$ in all the following analyses. The physics lesson here is that all six ground spaces (coupled with mass) encode similar information about the halo mass, which corresponds well with our intuition as luminosity is positively correlated with satellite galaxy mass and the three velocities are simply different components of the satellite velocity.

Third and most encouragingly, the values of $\text{Log}(M_{\text{sat}})$ corresponding to the minimum of the regressed function Δ_{OT} are all relatively close to the true value: the percentage difference is smaller than 1%. We therefore proceed to calculate the inferred likelihood using the BOLFI framework on this trial dataset of 500 halos; see the top two plots in Figure 4.8. Notice that here the regression function $\Delta_{\text{OT}}(M_{200})$ (first row; left) is the same as the subplot in the upper-left corner in Figure 4.7. BOLFI then gives the inferred likelihood function (first row; right) with its 95% confidence interval and the maximum likelihood estimate, i.e., $\text{Log}(M_{\text{sat}})_{\text{MLE}} = 11.917$ with the lower bound at 11.836 and the upper bound at 11.988. Although these inferred values are rather close to the ground truth, i.e., $\text{Log}(M_{\text{sat}})_{\text{truth}} = 12$, an unsettling issue is that the true value falls outside the 95% confidence interval. We now evoke active learning to add in more data points and see whether the inference performance will be improved.

The second row of Figure 4.8 shows the final regression plot $\Delta_{\text{OT}}(M_{200})$ and the inferred likelihood on the full dataset of 2500 halos. We run 20 iterations of active learning, where at each iteration 100 halos with the same value of M_{200} as determined by active learning are added to the dataset (shown as orange dots in the figure). Comparing to the result for the trial dataset, the inferred likelihood using the full dataset clearly has a narrower 95% confidence interval, suggesting the improved accuracy of the inference thanks to enhanced statistics. However, the MLE is now at $\text{Log}(M_{\text{sat}})_{\text{MLE}} = 11.908$, with the lower bound at 11.869 and the upper bound at 11.944. The ground truth $\text{Log}(M_{\text{sat}})_{\text{truth}} = 12$ is now even further away from both the MLE and the upper 95%

confidence interval bound. We therefore regard this inference task of halo mass to be semi-successful, in that a confidently constrained MLE can be obtained for the mass (in contrast to the inference of disc mass fraction) and yet the true value of M_{200} lies outside (specifically above) the 95% confidence interval of the inferred MLE.

Such low fits of M_{200} occur consistently no matter what OT distance we choose, as long as the ground space is picked from the observable features considered in this study for the halos, i.e., distance r, velocities v's, luminosity L, and log mass $\mathrm{Log}(M_{\mathrm{sat}})$. We have also tried simple summary statistics constructed out of the above features. Almost all of them uniformly give MLE fits lower than the actual $10^{12} M_{\odot}$. We therefore suspect that the features themselves are likely to be the cause of this "lower-fit problem". More generally, it is possible that using the dynamics of satellite galaxies alone may underestimate the virial mass of the host dark matter halo. A similar phenomenon has also been observed in [154]; see especially Section 8 for a detailed discussion.

Further studies are of course needed to gain better understanding of the "lower-fit problem" and more broadly for the various inference performances presented both in Section 4.3 and Section 4.4. Explanations in terms of the astrophysical underpinning and simulator upgrades are as important (if not more) as improvements on the sides of statistical framework developments. This would require deep domain expertises and we look forward to more collaborations with the astrophysics community for the further analysis of the two inference tasks here and for a broader application of the BOLFI+OT framework, with an eye towards inferring the particle nature of dark matter—especially its self-interaction strength.

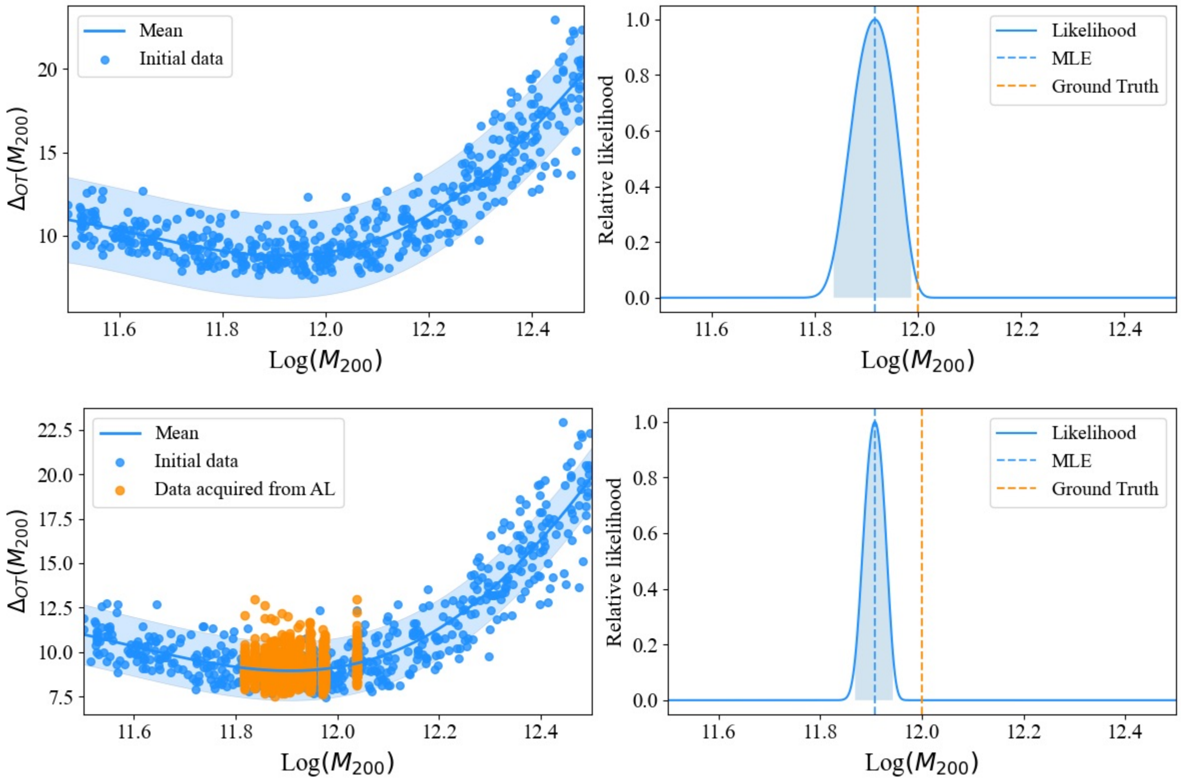


Figure 4.8: GP regression of $\Delta_{\rm OT}(M_{200})$ and the inferred likelihood function for M_{200} using the BOLFI framework on the trial dataset of 500 halos (first row) and on the full dataset of 2500 halos (second row) with the trial dataset serving as the initial dataset. The additional 2000 halos (orange dots) are obtained iteratively by active learning, where 20 iterations are run with each adding in 100 more halos at the same M_{200} value. For the likelihood functions, the ground truth value is given by the orange vertical line at $\text{Log}(M_{\rm sat})_{\rm truth} = 12$, whereas the inferred MLE is indicated as a blue vertical line with the associated 95% confidence interval showing as the blue shade. The MLE for the trial dataset is $\text{Log}(M_{\rm sat})_{\rm MLE} = 11.917$, and that for the full dataset is $\text{Log}(M_{\rm sat})_{\rm MLE} = 11.908$.

Chapter 5

Conclusion and Outlook

There certainly remains much more to explore at the interface between high energy physics and the theory of optimal transport. The present thesis has merely presented two illustrative examples, both still very much in their early stage of development. Yet we have already seen rather promising results, encouraging us both to further hone this powerful analysis tool and to dive deeper into the underlying theoretical connection between optimal transport and quantum field theory.

Let us recap what we have achieved in this thesis and offer some further ideas for future study. After briefly motivating our work in Chapter 1, we introduced in the next chapter two categories of optimal transport distances—balanced OT for distributions with equal total mass and unbalanced OT to generalize to the case of different total mass. In particular, we focused on two special OT distances, the balanced 2-Wasserstein metric and the unbalanced Hellinger-Kantorovich metric, and studied their linearization by locally approximating their weak Riemannian structure with a tangent space. In particular, we gave explicit forms of the respective logarithmic and exponential maps for W_2 and HK, as well as identifying a suitable notion of a Riemannian inner product for each. Data samples can thus be represented as vectors in the tangent space at a suitable reference measure where their Euclidean norm locally approximates the original metric. Such a linearization scheme is only available for the W_2 and HK distances, thanks to their unique geometric structure. Additionally, we put special emphasis on the discrete case, since in practical numerical applications one most certainly only encounters discrete distributions. We explicitly worked out every step to obtain the $\mathrm{LinW}_{2,\mathcal{R}}$ and $\mathrm{LinHK}_{\kappa,\mathcal{R}}$ pseudo-distances, which is the actual formalism deployed in the collider physics application.

Working with the local linearization and the corresponding embeddings allows us to take advantage of the Euclidean setting, such as a significantly reduced computational cost and a plethora of data analysis tools. As the same time, one can still enjoy the descriptive power of the original exact W_2 and HK metrics. The LOT framework developed here has significantly lowered the threshold for diverse applications of the theory of optimal transport to data sciences, including potentially many other scenarios in high energy physics.

Then in Chapter 3, we applied optimal transport in the context of collider physics, specifically for the task of jet tagging. The Euclidean embedding not only enables even a desktop computer to perform the OT calculation between $\mathcal{O}(10^5)$ collider events, but also makes a natural input to simple machine learning algorithms that require more than the pairwise distance between events. We demonstrated the value of the LOT framework—specifically $\mathrm{LinW}_{2,\mathcal{R}}$ and $\mathrm{LinHK}_{\kappa,\mathcal{R}}$—for jet tagging in a number of classification tasks, illustrating both the relative computational efficiency (compared to exact OT approaches) and interpretability (compared to deep neural networks) of our approach. The two main classifiers we employed in our study, kNN and SVM, when coupled with the LOT approximations, both achieve high performance on a level comparable to the exact OT approach and complex neural networks, while significantly outperforming the traditional N-subjettiness variable.

We then tackled the question—which is the best metric for the space of collider events, where we compared the performance on W *vs.* QCD jet tagging of $\mathrm{LinW}_{2,\mathcal{R}}$ and $\mathrm{LinHK}_{\kappa,\mathcal{R}}$ with a range of κ values. For optimized choices of κ, we found that $\mathrm{LinHK}_{\kappa,\mathcal{R}}$ matched or exceeded the same algorithms using $\mathrm{LinW}_{2,\mathcal{R}}$ or the original EMD distances in a fraction of the computing time. There is still considerable room to explore the interplay between the Hellinger-Kantorovich length scale parameter κ, the jet clustering radius, and the scale(s) associated with the choice of a reference measure. This subject has only been briefly touched upon and certainly worths future study. Furthermore, we presented a first study of the effects of pileup on optimal transport distances and found that boosted jet classification based on the LOT framework exhibited an encouraging degree of robustness against pileup contamination compared to the N-subjettiness shape observable. We also included a brief discussion of further upgrades to the optimal transport framework, with an eye mainly towards its application to event classification. Analysis on the full event level presents a number of additional challenges, all calling for innovative developments on the mathematical side. This serves as a good example to illustrate the impact of practical concerns on theoretical study, which, after all, is how the entire field of optimal transport was founded in the first place.

By equipping the space of jets with a metric, the theory of optimal transport offers a new perspective on the traditional problems of collider physics, from unifying the panoply of collider observables to enabling the use of interpretable distance-based machine learning algorithms. The computational speedup offered by our LOT approximation should make it possible to apply optimal transport methods more broadly in analyzing both simulated and actual collider data. Although we have mostly focused on boosted jet classification as an initial application to collider physics, the flexible LOT framework should be generally well-suited to an array of applications beyond supervised learning, including clustering and anomaly detection. More broadly, the Riemannian event manifold itself obtained with either the 2-Wasserstein or the Hellinger-Kantorovich distance is likely to have interesting properties and may reveal further hidden structure in the space of collider events.

The same properties that make optimal transport an ideal tool for studying jet substructure at colliders also make it potentially suitable to compare dark matter halo substructure between simulation and observation. Chapter 4 offered the first systematic use of optimal transport in comparing simulated dark matter subhalo distributions with the aim of inferring the underlying halo property. Although the ultimate dream is to investigate the effects of dark matter self-interactions and its other microscopic properties, here we set as our immediate goal the inference of disc mass fraction and halo mass.

As the physics problem belongs to the realm of simulation-based inference, we rephrased it in the more general statistical language and introduced the Bayesian Optimization for Likelihood-free Inference framework, where active learning is employed for increased sampling efficiency in face of expensive simulators. Our innovation on the framework development side lies in the replacement of the usual handpicked summary statistics to quantify similarity between halos with a more automated and sophisticated notion of distance based on optimal transport.

Here the representation of halos as distributions does not come as naturally as in the case of jets. The choice of a suitable ground space not only critically encodes the underlying physics but is also task specific. We have seen that the same halo representation resulted in vastly different outcomes when the inference is for the disc mass fraction or for the halo mass. For f_m, all ground space choices failed equally miserably, suggesting that the information contained in the current ground space(s) is not enough, or even worse entirely off the point. On the other hand, the inference of halo mass is much more successful. Here all six choices of the ground space give similar performance, indicating in a different way that they encode more or less the same physics about halos. The obvious next step is to incorporate a larger variety of observable features of halo substructure, in addition to the radii, velocities, magnitudes, and etc, currently being used. That would in particular require further upgrades from the astrophysics side, as these features should be either directly measurable or easily deducible from measurable quantities in order to facilitate the envisioned simulation-to-observation comparison.

The current mixed results are in some sense more encouraging than a "happy ending" everyone initially was wishing for. For one thing, it compels us to examine more closely the astrophysical reasons behind the success or the failure of certain choices of the ground space. It also highlights the importance of halo-to-halo variance, which may call for a more versatile regression model than the Gaussian process currently implemented in the BOLFI framework. Furthermore, if unfortunately optimal transport ends up failing to provide a satisfying distance between dark matter halos, it would indeed make the success of OT in collider physics an even stronger case and hint at deeper theoretical connection between optimal transport and jet physics—a topic that has been studied in the pioneering work [13]. But let us hope for the best, in which case our current effort is laying the groundwork for future studies of dark matter self-interactions wherein subhalo distributions can be compared as a function of dark matter self-interaction strength using the same method. Of course, there is still a long way to go, and the first step is to resolve the aforementioned "lower fit problem" in the inference of halo mass. More work needs to be done, both on the framework development side and on the side of astrophysical simulations.

5.1 Final Thoughts on Cross-disciplinary Research

As mentioned at the beginning of this thesis, many scientific fields today are facing an explosion of data thanks to expansion and improvement of experimental pursuits, and are therefore similarly in need of advanced statistical methods to help extract useful scientific knowledge from the rich information collected. In many cases, though the underlying scientific focuses may be vastly different, the data themselves nonetheless often possess a high degree of similarity that would allow for a unified treatment. Hence, the development of versatile analysis frameworks naturally transcends the traditional disciplinary boundaries and calls for more conversations and collaborations between different research fields.

Artificial Intelligence promises such a powerful set of tools based mainly on deep neural networks but also including a variety of other methods. Traditionally, AI methods are tested and deployed on industry-prepared datasets which may hold more commercial values, such as images for computer vision and texts for natural language processing. Yet increasingly, researchers in fundamental sciences are realizing the great benefits AI may bring to their individual fields and we now see blooming endeavors almost everywhere to incorporate AI into their standard analysis pipeline. In order not to repeat the same effort of introducing and developing AI frameworks over and over again under different contexts, the best way is for researchers across a range of scientific fields to talk to each other and work together to solve the same underlying data problem despite the apparent difference in the scientific questions they pursue. The present thesis gives one example of how the early success of optimal transport in fields such as computer vision, economics, and medical imaging translates to surprising benefits for high energy physics research. In turn, we expect that the statistical frameworks developed here and the deeper theoretical analysis will have positive impacts on other fields and would therefore like to encourage everyone to think more about its potential usages elsewhere.

As every research field today is blessed with essentially the same challenge, more cross-disciplinary dialogues should be encouraged to foster collaborations among different domain experts. In particular, we believe that high energy physics has a unique advantage with lots to offer and will continue to contribute to this big data and AI revolution that is unfolding both across all research fields and on the larger societal scale.

Bibliography

[1] HEP ML Community, "A Living Review of Machine Learning for Particle Physics."

[2] B. P. Roe, H.-J. Yang, J. Zhu, Y. Liu, I. Stancu, and G. McGregor, *Boosted Decision Trees as an Alternative to Artificial Neural Networks for Particle Identification*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **543** (May, 2005) 577–584.

[3] H.-J. Yang, B. P. Roe, and J. Zhu, *Studies of Boosted Decision Trees for MiniBooNE Particle Identification*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **555** (Dec, 2005) 370–385.

[4] P. T. Komiske, E. M. Metodiev, and J. Thaler, *Metric Space of Collider Events*, *Phys. Rev. Lett.* **123** (2019), no. 4 041801, [arXiv:1902.0234].

[5] M. Thorpe, S. Park, S. Kolouri, G. K. Rohde, and D. Slepčev, *A Transportation L^p Distance for Signal Analysis*, *Journal of mathematical imaging and vision* **59** (2017), no. 2 187–210.

[6] O. Pele and M. Werman, *A Linear Time Histogram Metric for Improved Sift Matching*, in *European conference on computer vision*, pp. 495–508, Springer, 2008.

[7] O. Pele and M. Werman, *Fast and Robust Earth Mover's Distances*, in *2009 IEEE 12th International Conference on Computer Vision*, pp. 460–467, IEEE, 2009.

[8] Y. Rubner, C. Tomasi, and L. J. Guibas, *The Earth Mover's Distance as a Metric for Image Retrieval*, *International journal of computer vision* **40** (2000), no. 2 99–121.

[9] W. Wang, J. A. Ozolek, D. Slepčev, A. B. Lee, C. Chen, and G. K. Rohde, *An Optimal Transportation Approach for Nuclear Structure-based Pathology*, *IEEE transactions on medical imaging* **30** (2010), no. 3 621–631.

[10] J. Delon, *Midway Image Equalization*, *Journal of Mathematical Imaging and Vision* **21** (2004), no. 2 119–134.

[11] Y. Rubner, C. Tomasi, and L. J. Guibas, *A Metric for Distributions with Applications to Image Databases*, in *Proceedings of the Sixth International Conference on Computer Vision*, ICCV '98, (USA), p. 59, IEEE Computer Society, 1998.

[12] S. Peleg, M. Werman, and H. Rom, *A Unified Approach to the Change of Resolution: Space and Gray-level*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11** (1989), no. 7 739–742.

[13] P. T. Komiske, E. M. Metodiev, and J. Thaler, *The Hidden Geometry of Particle Collisions*, *JHEP* **07** (Jul, 2020) 006, [arXiv:2004.0415].

[14] P. T. Komiske, R. Mastandrea, E. M. Metodiev, P. Naik, and J. Thaler, *Exploring the Space of Jets with CMS Open Data*, *Phys. Rev. D* **101** (2020), no. 3 034009, [arXiv:1908.0854].

[15] C. Cesarotti and J. Thaler, *A Robust Measure of Event Isotropy at Colliders*, arXiv:2004.0612.

[16] M. C. Romao, N. Castro, J. Milhano, R. Pedro, and T. Vale, *Use of a Generalized Energy Mover's Distance in the Search for Rare Phenomena at Colliders*, arXiv:2004.0936.

[17] S. E. Park, P. Harris, and B. Ostdiek, *Neural Embedding: Learning the Embedding of the Manifold of Physics Data*, 2022.

[18] C. Pollard and P. Windischhofer, *Transport Away Your Problems: Calibrating Stochastic Simulations with Optimal Transport*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **1027** (Mar, 2022) 166119.

[19] A. Davis, T. Menzo, A. Youssef, and J. Zupan, *Earth mover's distance as a measure of CP violation*, 2023.

[20] D. Ba, A. S. Dogra, R. Gambhir, A. Tasissa, and J. Thaler, *SHAPER: Can You Hear the Shape of a Jet?*, 2023.

[21] S. Alipour-fard, P. T. Komiske, E. M. Metodiev, and J. Thaler, *Pileup and Infrared Radiation Annihilation (PIRANHA): A Paradigm for Continuous Jet Grooming*, 2023.

[22] A. J. Larkoski and J. Thaler, *A Spectral Metric for Collider Geometry*, 2023.

[23] T. Cheng, J.-F. Arguin, J. Leissner-Martin, J. Pilette, and T. Golling, *Variational Autoencoders for Anomalous Jet Tagging*, arXiv:2007.0185.

[24] S. Tsan, R. Kansal, A. Aportela, D. Diaz, J. Duarte, S. Krishna, F. Mokhtar, J.-R. Vlimant, and M. Pierini, *Particle Graph Autoencoders and Differentiable, Learned Energy Mover's Distance*, 2021.

[25] J. H. Collins, *An Exploration of Learnt Representations of W Jets*, 9, 2021. arXiv:2109.1091.

[26] J. N. Howard, S. Mandt, D. Whiteson, and Y. Yang, *Learning to Simulate High Energy Particle Collisions from Unlabeled Data*, *Sci. Rep.* **12** (2022) 7567, [arXiv:2101.0894].

[27] L. Gouskos, F. Iemmi, S. Liechti, B. Maier, V. Mikuni, and H. Qu, *Optimal Transport for a Global Event Description at High-intensity Hadron Colliders*, 2023.

[28] A. Mullin, H. Pacey, M. Parker, M. White, and S. Williams, *Does SUSY have friends? A new approach for LHC event analysis*, arXiv:1912.1062.

[29] J. Cotler and S. Rezchikov, *Renormalization Group Flow as Optimal Transport*, 2023.

[30] D. S. Berman and M. S. Klinger, *The Inverse of Exact Renormalization Group Flows as Statistical Inference*, 2022.

[31] D. S. Berman, M. S. Klinger, and A. G. Stapleton, *Bayesian Renormalization*, 2023.

[32] G. B. De Luca, N. De Ponti, A. Mondino, and A. Tomasiello, *Gravity from Thermodynamics: Optimal Transport and Negative Effective Dimensions*, arXiv:2212.0251.

[33] G. B. D. Luca, N. D. Ponti, A. Mondino, and A. Tomasiello, *Cheeger Bounds on Spin-two Fields*, *Journal of High Energy Physics* **2021** (Dec, 2021).

[34] T. Cai, J. Cheng, B. Schmitzer, and M. Thorpe, *The Linearized Hellinger–Kantorovich Distance*, *SIAM Journal on Imaging Sciences* **15** (2022), no. 1 45–83, [https://doi.org/10.1137/21M1400080].

[35] T. Cai, J. Cheng, N. Craig, and K. Craig, *Linearized Optimal Transport for Collider Events*, *Physical Review D* **102** (Dec, 2020).

[36] T. Cai, J. Cheng, K. Craig, and N. Craig, *Which Metric on the Space of Collider Events?*, *Physical Review D* **105** (Apr, 2022).

[37] G. Peyré, M. Cuturi, *et. al.*, *Computational Optimal Transport: With Applications to Data Science*, *Foundations and Trends in Machine Learning* **11** (2019), no. 5-6 355–607.

[38] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows: In Metric Spaces and In the space of Probability Measures.* Springer Science & Business Media, 2008.

[39] F. Santambrogio, *Optimal Transport for Applied Mathematicians*, *Birkäuser, NY* **55** (2015), no. 58-63 94.

[40] C. Villani, *Topics in Optimal Transportation.* No. 58. American Mathematical Soc., 2003.

[41] C. Villani, *Optimal Transport: Old and New.* A Series of Comprehensive Studies in Mathematics. Springer Science & Business Media, 2008.

[42] S. Kolouri, S. R. Park, M. Thorpe, D. Slepčev, and G. K. Rohde, *Optimal Mass Transport: Signal Processing and Machine Learning Applications*, *IEEE Signal Processing Magazine* **34** (2017), no. 4 43–59.

[43] L. C. Torres, L. M. Pereira, and M. H. Amini, *A Survey on Optimal Transport for Machine Learning: Theory and Applications*, 2021.

[44] N. Gigli, *On the Inverse Implication of Brenier-McCann Theorems and the Structure of* $(P_2(M), W_2)$, *Methods Appl. Anal.* **18** (2011), no. 2 127–158.

[45] R. J. McCann, *Existence and Uniqueness of Monotone Measure-preserving Maps*, *Duke Math. J.* **80** (1995), no. 2 309–323.

[46] J.-D. Benamou and Y. Brenier, *A Computational Fluid Mechanics Solution to the Monge–Kantorovich Mass Transfer Problem*, *Numerische Mathematik* **84** (2000), no. 3 375–393.

[47] Y. Brenier, *Polar Factorization and Monotone Rearrangement of Vector-valued Functions*, *Comm. Pure Appl. Math.* **44** (1991), no. 4 375–417.

[48] A. Figalli, *The Optimal Partial Transport Problem*, *Archive for rational mechanics and analysis* **195** (2010), no. 2 533–560.

[49] B. Piccoli and F. Rossi, *Generalized Wasserstein Distance and Its Application to Transport Equations with Source*, *Archive for Rational Mechanics and Analysis* **211** (2014), no. 1 335–358.

[50] B. Piccoli and F. Rossi, *On Properties of the Generalized Wasserstein Distance*, *Archive for Rational Mechanics and Analysis* **222** (2016), no. 3 1339–1365.

[51] M. Liero, A. Mielke, and G. Savaré, *Optimal Transport in Competition with Reaction: The Hellinger–Kantorovich Distance and Geodesic Curves*, *SIAM Journal on Mathematical Analysis* **48** (2016), no. 4 2869–2911.

[52] M. Liero, A. Mielke, and G. Savaré, *Optimal Entropy-transport Problems and a New Hellinger–Kantorovich Distance between Positive Measures*, *Inventiones mathematicae* **211** (2018), no. 3 969–1117.

[53] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, *An Interpolating Distance between Optimal Transport and Fisher–Rao Metrics*, *Found. Comp. Math.* **18** (2018), no. 1 1–44.

[54] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, *Scaling Algorithms for Unbalanced Optimal Transport Problems*, *Mathematics of Computation* **87** (2018), no. 314 2563–2609.

[55] S. Kondratyev, L. Monsaingeon, D. Vorotnikov, *et. al.*, *A New Optimal Transport Distance on the Space of Finite Radon Measures*, *Advances in Differential Equations* **21** (2016), no. 11/12 1117–1164.

[56] W. Gangbo, W. Li, S. Osher, and M. Puthawala, *Unnormalized Optimal Transport*, *Journal of Computational Physics* **399** (2019) 108940.

[57] E. K. Ryu, W. Li, P. Yin, and S. Osher, *Unbalanced and Partial L_1 Monge–Kantorovich Problem: A Scalable Parallel First-Order Method*, *Journal of Scientific Computing* **75** (2018), no. 3 1596–1613.

[58] L. A. Caffarelli and R. J. McCann, *Free Boundaries in Optimal Transport and Monge-Ampère Obstacle Problems*, *Annals of Math.* **171** (2010), no. 2 673–730.

[59] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard, *Unbalanced Optimal Transport: Dynamic and Kantorovich Formulations*, *J. Funct. Anal.* **274** (2018), no. 11 3090–3123.

[60] J. Dolbeault, B. Nazaret, and G. Savaré, *A New Class of Transport Distances between Measures*, *Calc. Var. Partial Differential Equations* **34** (2009), no. 2 193–231.

[61] B. Schmitzer and B. Wirth, *A Framework for Wasserstein-1-Type Metrics*, *Journal of Convex Analysis* **26** (2019), no. 2 353–396.

[62] T. T. Georgiou, J. Karlsson, and M. S. Takyar, *Metrics for Power Spectra: An Axiomatic Approach*, *IEEE Transactions on Signal Processing* **57** (2008), no. 3 859–867.

[63] V. Laschos and A. Mielke, *Geometric Properties of Cones with Applications on the Hellinger–Kantorovich Space, and a New Distance on the Space of Probability Measures*, *J. Funct. Anal.* **276** (2019), no. 11 3529–3576.

[64] M. Agueh and G. Carlier, *Barycenters in the Wasserstein Space*, *SIAM J. Math. Anal.* **43** (2011), no. 2 904–924.

[65] G. Peyré and M. Cuturi, *Computational Optimal Transport*, *Foundations and Trends in Machine Learning* **11** (2019), no. 5–6 355–607.

[66] J. Altschuler, J. Niles-Weed, and P. Rigollet, *Near-linear Time Approximation Algorithms for Optimal Transport via Sinkhorn Iteration*, in *Advances in Neural Information Processing Systems*, pp. 1964–1974, 2017.

[67] D. P. Bertsekas, *A New Algorithm for the Assignment Problem*, *Mathematical Programming* **21** (1981), no. 1 152–171.

[68] D. P. Bertsekas and J. Eckstein, *Dual Coordinate Step Methods for Linear Network Flow Problems*, *Mathematical Programming* **42** (1988), no. 1-3 203–243.

[69] M. Cuturi, *Sinkhorn Distances: Lightspeed Computation of Optimal Transport*, in *Advances in Neural Information Processing Systems*, pp. 2292–2300, 2013.

[70] W. Wang, D. Slepčev, S. Basu, J. A. Ozolek, and G. K. Rohde, *A Linear Optimal Transportation Framework for Quantifying and Visualizing Variations in Sets of Images*, *International journal of computer vision* **101** (2013), no. 2 254–269.

[71] A. Delalande and Q. Merigot, *Quantitative Stability of Optimal Transport Maps under Variations of the Target Measure*, *preprint arXiv:2103.05934* (2021).

[72] S. R. Park and M. Thorpe, *Representing and Learning High Dimensional Data with the Optimal Transport Map from a Probabilistic Viewpoint*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7864–7872, 2018.

[73] S. Kolouri and G. K. Rohde, *Transport-Based Single Frame Super Resolution of Very Low Resolution Face Images*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4876–4884, 2015.

[74] J. A. Ozolek, A. B. Tosun, W. Wang, C. Chen, S. Kolouri, S. Basu, H. Huang, and G. K. Rohde, *Accurate Diagnosis of Thyroid Follicular Lesions from Nuclear Morphology Using Supervised Learning*, *Medical image analysis* **18** (2014), no. 5 772–780.

[75] L. Ambrosio and N. Gigli, *A User's Guide to Optimal Transport*, in *Modelling and Optimisation of Flows on Networks*, vol. 2062 of *Lect. Not. Math.*, pp. 1–155. Springer, 2013.

[76] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows in Metric Spaces and in the Space of Probability Measures.* Lectures in Mathematics. Birkhäuser Boston, 2005.

[77] Q. Mérigot, A. Delalande, and F. Chazal, *Quantitative Stability of Optimal Transport Maps and Linearization of the 2-wasserstein Space*, in *International Conference on Artificial Intelligence and Statistics*, pp. 3186–3196, 2020.

[78] G. Friesecke, D. Matthes, and B. Schmitzer, *Barycenters for the Hellinger–Kantorovich Distance Over $\mathbb{R}^d$*, *SIAM Journal on Mathematical Analysis* **53** (Jan, 2021) 62–110.

[79] R. J. Berman, *Convergence Rates for Discretized Monge–Ampère Equations and Quantitative Stability of Optimal Transport*, *Foundations of Computational Mathematics* (2020).

[80] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré, *Iterative Bregman Projections for Regularized Transportation Problems*, *SIAM Journal on Scientific Computing* **37** (2015), no. 2 A1111–A1138.

[81] N.-P. Chung and M.-N. Phung, *Barycenters in the Hellinger-Kantorovich Space*, 2020.

[82] B. Schmitzer, *Stabilized Sparse Scaling Algorithms for Entropy Regularized Transport Problems*, 2019.

[83] G. F. Sterman and S. Weinberg, *Jets from Quantum Chromodynamics*, *Phys. Rev. Lett.* **39** (1977) 1436.

[84] M. Cacciari, G. P. Salam, and G. Soyez, *FastJet User Manual*, *The European Physical Journal C* **72** (Mar, 2012).

[85] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N-subjettiness*, *Journal of High Energy Physics* **2011** (Mar, 2011).

[86] J. Cogan, M. Kagan, E. Strauss, and A. Schwarztman, *Jet-Images: Computer Vision Inspired Techniques for Jet Tagging*, *Journal of High Energy Physics* **2015** (Feb, 2015).

[87] P. T. Komiske, E. M. Metodiev, and J. Thaler, *Energy Flow Networks: Deep Sets for Particle Jets*, *Journal of High Energy Physics* **2019** (Jan, 2019).

[88] A. J. Larkoski, I. Moult, and B. Nachman, *Jet Substructure at the Large Hadron Collider: A Review of Recent Advances in Theory and Machine Learning*, *Physics Reports* **841** (Jan, 2020) 1–63.

[89] P. Calafiura, D. Rousseau, and K. Terao, *Artificial Intelligence for High Energy Physics.* WORLD SCIENTIFIC, 2022.

[90] P. T. Komiske, E. M. Metodiev, and J. Thaler, *Energy Flow Polynomials: A Complete Linear Basis for Jet Substructure*, *Journal of High Energy Physics* **2018** (Apr, 2018).

[91] F. V. Tkachov, *Measuring Multijet Structure of Hadronic Energy Flow or What IS A jet?*, *International Journal of Modern Physics A* **12** (Dec, 1997) 5411–5529.

[92] R. Flamary and N. Courty, *POT Python Optimal Transport Library*, 2017.

[93] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, *Scikit-learn: Machine Learning in Python*, *Journal of Machine Learning Research* **12** (2011) 2825–2830.

[94] R. A. Fisher, *The use of Multiple Measurements in Taxonomic Problems*, *Annals of Eugenics* **7** (1936), no. 2 179–188, [https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1936.tb02137.x].

[95] T. Cover and P. Hart, *Nearest Neighbor Pattern Classification*, *IEEE Transactions on Information Theory* **13** (1967), no. 1 21–27.

[96] C. Cortes and V. N. Vapnik, *Support-Vector Networks*, *Machine Learning* **20** (1995), no. 3 273–297.

[97] L. Kaufman and P. Rousseeuw, *Clustering by Means of Medoids*, in *Statistical Data Analysis Based on the L_1–Norm and Related Methods* (Y. Dodge, ed.), pp. 405–416. Springer, 1987.

[98] A. V. Novikov, *PyClustering: Data Mining Library*, *Journal of Open Source Software* **4** (2019), no. 36 1230.

[99] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, *The Automated Computation of Tree-level and Next-to-leading Order Differential Cross Sections, and Their Matching to Parton Shower Simulations*, *Journal of High Energy Physics* **2014** (Jul, 2014).

[100] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, *An Introduction to PYTHIA 8.2*, *Computer Physics Communications* **191** (Jun, 2015) 159–177.

[101] P. T. Komiske, E. M. Metodiev, and J. Thaler, *Cutting Multiparticle Correlators Down to Size*, *Phys. Rev. D* **101** (2020), no. 3 036019, [arXiv:1911.0449].

[102] J. Thaler and K. Van Tilburg, *Maximizing Boosted Top Identification by Minimizing N-subjettiness*, *Journal of High Energy Physics* **2012** (Feb, 2012).

[103] **CMS** Collaboration, *Jet Energy Scale and Resolution in the CMS Experiment in pp Collisions at 8 TeV*, *Journal of Instrumentation* **12** (Feb, 2017) P02014–P02014.

[104] **ATLAS** Collaboration, *Jet Energy Scale Measurements and Their Systematic Uncertainties in proton-proton Collisions at $\sqrt{s} = 13$ TeV with the ATLAS Detector*, *Physical Review D* **96** (Oct, 2017).

[105] **CMS** Collaboration, *Pileup Jet Identification*, CMS-PAS-JME-13-005, 2013.

[106] **ATLAS** Collaboration, *Performance of the ATLAS Inner Detector Track and Vertex Reconstruction in the High Pile-Up LHC Environment*, ATLAS-CONF-2012-042, 3, 2012.

[107] **ATLAS** Collaboration, G. Aad *et. al.*, *Performance of Pile-up Mitigation Techniques for Jets in pp Collisions at $\sqrt{s} = 8$ TeV Using the ATLAS Detector*, *Eur. Phys. J. C* **76** (2016), no. 11 581, [arXiv:1510.0382].

[108] G. P. Salam, *Towards Jetography*, *Eur. Phys. J. C* **67** (2010) 637–686, [arXiv:0906.1833].

[109] G. Soyez, *Pileup Mitigation at the LHC: A Theorist's View*, .

[110] P. Berta, M. Spousta, D. W. Miller, and R. Leitner, *Particle-level Pileup Subtraction for Jets and Jet Shapes*, *JHEP* **06** (2014) 092, [arXiv:1403.3108].

[111] D. Krohn, M. D. Schwartz, M. Low, and L.-T. Wang, *Jet Cleansing: Pileup Removal at High Luminosity*, *Phys. Rev. D* **90** (2014), no. 6 065020, [arXiv:1309.4777].

[112] D. Bertolini, P. Harris, M. Low, and N. Tran, *Pileup Per Particle Identification*, *JHEP* **10** (2014) 059, [arXiv:1407.6013].

[113] M. Cacciari, G. P. Salam, and G. Soyez, *SoftKiller, A Particle-level Pileup Removal Method*, *Eur. Phys. J. C* **75** (2015), no. 2 59, [arXiv:1407.0408].

[114] P. T. Komiske, E. M. Metodiev, B. Nachman, and M. D. Schwartz, *Pileup Mitigation with Machine Learning (PUMML)*, *JHEP* **12** (2017) 051, [arXiv:1707.0860].

[115] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, *Techniques for Improved Heavy Particle Searches with Jet Substructure*, *Phys. Rev. D* **80** (2009) 051501, [arXiv:0903.5081].

[116] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, *Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches*, *Phys. Rev. D* **81** (2010) 094023, [arXiv:0912.0033].

[117] **CMS** Collaboration, *Identifying Hadronically Decaying Vector Bosons Merged into a Single Jet*, CMS-PAS-JME-13-006, 2013.

[118] D. Akerib and *et al.*, *The LUX-ZEPLIN (LZ) Experiment*, *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **953** (2020) 163047.

[119] J. Aalbers and *et al.*, *First Dark Matter Search Results from the LUX-ZEPLIN (LZ) Experiment*, 2022.

[120] J. Aalbers and *et al.*, *Background Determination for the LUX-ZEPLIN (LZ) Dark Matter Experiment*, 2022.

[121] J. M. Gaskins, *A Review of Indirect Searches for Particle Dark Matter*, *Contemporary Physics* **57** (Jun, 2016) 496–525.

[122] R. K. Leane, *Indirect Detection of Dark Matter in the Galaxy*, 2020.

[123] R. A. Flores and J. R. Primack, *Observational and Theoretical Constraints on Singular Dark Matter Halos*, *arXiv preprint astro-ph/9402004* (1994).

[124] B. Moore, *Evidence Against Dissipation-less Dark Matter from Observations of Galaxy Haloes*, *Nature* **370** (1994) 629.

[125] A. B. Newman, T. Treu, R. S. Ellis, D. J. Sand, C. Nipoti, J. Richard, and E. Jullo, *The Density Profiles of Massive, Relaxed Galaxy Clusters. I. The Total Density Over Three Decades in Radius*, *The Astrophysical Journal* **765** (Feb, 2013) 24.

[126] A. B. Newman, T. Treu, R. S. Ellis, and D. J. Sand, *The Density Profiles of Massive, Relaxed Galaxy Clusters. II. Separating Luminous and Dark Matter in Cluster Cores*, *The Astrophysical Journal* **765** (Feb, 2013) 25.

[127] A. B. Newman, R. S. Ellis, and T. Treu, *Luminous and Dark Matter Profiles from Galaxies to Clusters: Bridging the Gap with Group-scale Lenses*, 2015.

[128] A. Klypin, A. V. Kravtsov, O. Valenzuela, and F. Prada, *Where Are the Missing Galactic Satellites?*, *The Astrophysical Journal* **522** (Sep, 1999) 82–92.

[129] B. Moore, S. Ghigna, F. Governato, G. Lake, T. Quinn, J. Stadel, and P. Tozzi, *Dark Matter Substructure within Galactic Halos*, *The Astrophysical Journal* **524** (Oct, 1999) L19–L22.

[130] L. Gao, S. D. M. White, A. Jenkins, F. Stoehr, and V. Springel, *The Subhalo Populations of ΛCDM Dark Haloes*, *Monthly Notices of the Royal Astronomical Society* **355** (Dec, 2004) 819–834.

[131] J. Diemand, M. Kuhlen, and P. Madau, *Formation and Evolution of Galaxy Dark Matter Halos and Their Substructure*, *The Astrophysical Journal* **667** (Oct, 2007) 859–877.

[132] V. Springel, J. Wang, M. Vogelsberger, A. Ludlow, A. Jenkins, A. Helmi, J. F. Navarro, C. S. Frenk, and S. D. M. White, *The Aquarius Project: The Subhaloes of Galactic Haloes*, *Monthly Notices of the Royal Astronomical Society* **391** (Dec, 2008) 1685–1711.

[133] M. Boylan-Kolchin, J. S. Bullock, and M. Kaplinghat, *Too big to fail? The Puzzling Darkness of Massive Milky Way Subhaloes*, *Monthly Notices of the Royal Astronomical Society: Letters* **415** (Jun, 2011) L40–L44.

[134] D. N. Spergel and P. J. Steinhardt, *Observational Evidence for Self-Interacting Cold Dark Matter*, *Physical Review Letters* **84** (Apr, 2000) 3760–3763.

[135] J. S. Bullock and M. Boylan-Kolchin, *Small-Scale Challenges to the ΛCDM Paradigm*, *Annual Review of Astronomy and Astrophysics* **55** (Aug, 2017) 343–387.

[136] J. F. Navarro, C. S. Frenk, and S. D. M. White, *A Universal Density Profile from Hierarchical Clustering*, *The Astrophysical Journal* **490** (Dec, 1997) 493–508.

[137] R. H. Wechsler, J. S. Bullock, J. R. Primack, A. V. Kravtsov, and A. Dekel, *Concentrations of Dark Halos from Their Assembly Histories*, *The Astrophysical Journal* **568** (Mar, 2002) 52.

[138] F. Jiang, A. Dekel, J. Freundlich, F. C. van den Bosch, S. B. Green, P. F. Hopkins, A. Benson, and X. Du, *SatGen: A Semi-analytical Satellite Galaxy Generator – I. The Model and Its Application to Local-Group Satellite Statistics*, *Monthly Notices of the Royal Astronomical Society* **502** (Jan, 2021) 621–641.

[139] S. B. Green, F. C. van den Bosch, and F. Jiang, *SatGen – II. Assessing the Impact of a Disc Potential on Subhalo Populations*, *Monthly Notices of the Royal Astronomical Society* **509** (Oct, 2021) 2624–2636.

[140] H. Parkinson, S. Cole, and J. Helly, *Generating Dark Matter Halo Merger Trees*, *Monthly Notices of the Royal Astronomical Society* **383** (Dec, 2007) 557–564.

[141] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, *Normalizing Flows for Probabilistic Modeling and Inference*, *Journal of Machine Learning Research* **22** (2021), no. 57 1–64.

[142] D. B. Rubin, *Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician*, *The Annals of Statistics* **12** (1984), no. 4 1151 – 1172.

[143] M. U. Gutmann and J. Corander, *Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models*, 2015.

[144] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning.* The MIT Press, 11, 2005.

[145] D. D. Cox and S. John, *A Statistical Method for Global Optimization*, *[Proceedings] 1992 IEEE International Conference on Systems, Man, and Cybernetics* (1992) 1241–1246 vol.2.

[146] M. Rosenblatt, *Remarks on Some Nonparametric Estimates of a Density Function*, *Annals of Mathematical Statistics* **27** (1956) 832–837.

[147] M. P. Wand and M. C. Jones, *Kernel Smoothing.* CRC Press, 1994.

[148] R. J. Rossi, *Mathematical Statistics: An Introduction to Likelihood Based Inference.* Wiley, 2018.

[149] L. Wang, A. A. Dutton, G. S. Stinson, A. V. Macciò, C. Penzo, X. Kang, B. W. Keller, and J. Wadsley, *NIHAO project – I. Reproducing the Inefficiency of Galaxy Formation Across Cosmic Time with a Large Sample of Cosmological Hydrodynamical Simulations*, *Monthly Notices of the Royal Astronomical Society* **454** (09, 2015) 83–94, [https://academic.oup.com/mnras/article-pdf/454/1/83/3914553/stv1937.pdf].

[150] Z.-Z. Li, D.-H. Zhao, Y. P. Jing, J. Han, and F.-Y. Dong, *Orbital Distribution of Infalling Satellite Halos Across Cosmic Time*, *Astrophys. J.* **905** (2020), no. 2 177, [arXiv:2008.0571].

[151] S. Green, F. Jiang, Z. Li, and D. Folsom, ***SatGen Github***, 2020.

[152] E. D'Onghia, V. Springel, L. Hernquist, and D. Keres, *Substructure Depletion In the Milky Way Halo by The Disk*, *The Astrophysical Journal* **709** (Jan, 2010) 1138.

[153] S. Garrison-Kimmel, A. Wetzel, J. S. Bullock, P. F. Hopkins, M. Boylan-Kolchin, C.-A. Faucher-Giguère, D. Kereš, E. Quataert, R. E. Sanderson, A. S. Graus, and T. Kelley, *Not So Lumpy After All: Modelling the Depletion of Dark Matter Subhaloes by Milky Way-like Galaxies*, *Monthly Notices of the Royal Astronomical Society* **471** (07, 2017) 1709–1727, [https://academic.oup.com/mnras/article-pdf/471/2/1709/19407370/stx1710.pdf].

[154] W. Wang, J. Han, M. Cautun, Z. Li, and M. N. Ishigaki, *The Mass of Our Milky Way*, *Science China Physics, Mechanics &amp Astronomy* **63** (May, 2020).