On anomaly detection in particle accelerators

by

Nilai Manish Sarda

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Author .....
Department of Electrical Engineering and Computer Science
May 12, 2020

Certified by .....
Justin M. Solomon
Associate Professor
Thesis Supervisor

Accepted by .....
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

On anomaly detection in particle accelerators

by

Nilai Manish Sarda

Submitted to the Department of Electrical Engineering and Computer Science
on May 12, 2020, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

In this thesis, we explore concepts related to anomaly detection at particle accelerators. In an extremely high-energy proton-proton collision, sprays of subatomic particles are generated and destroyed at extremely short timescales. Detecting anomalies in these event signatures is a crucial step in experimentally verifying new theories of physics beyond the Standard Model.

Towards this end, we analyze the geometric structure of event signatures, which can be represented as discrete distributions of energy on the surface of a cylinder. Therefore, we approach the problem from the framework of optimal transportation. The optimal transport distance is the map between two measures which minimizes the total work required to transform one measure into another.

First, we provide theoretical improvements in learning with transport-type distances. We show that minimizing the transport distance also leads to a coreset with respect to a broad class of functions, by showing a bound on the quadrature error of a Monte Carlo integration. In addition, we develop an unbiased estimator for a Gaussian kernel based on the sliced Wasserstein distance, which is based on the one-dimensional version of optimal transport.

Next, we use these kernels within the framework of discriminative anomaly detection. The methods we consider apply transport distance-based kernels to classify anomalies on an event-by-event basis. We apply these techniques to two datasets, one from the field of particle physics and one inspired by biology. This comparison allows us to argue empirically which models are most effective for which types of anomalies.

Finally, we move to the generative setting, and build a topic model based on the dijet factorization theorem to perform anomaly detection and quark/gluon discrimination. This approach leverages the fact that each jet in a dijet pair is statistically independent, and uses matrix factorization to disentangle the component distributions.

Thesis Supervisor: Justin M. Solomon
Title: Associate Professor

Acknowledgments

Finishing this thesis in the middle of a pandemic has been a challenging endeavour, and I would not have been able to complete it without the love and support of my community at MIT. Throughout my time here, I've been lucky enough to meet and befriend a variety of wonderful, wonderful people. They have made the past four years the best experience of my life. While this section is far too short to fully acknowledge them in the manner they deserve, I would like to briefly thank the following people:

- Justin Solomon for advising me so patiently through a SuperUROP and this M.Eng. This work would not have been possible without his constant advice and guidance.

- Ed Chien, Sebastian Claici, Aude Genevay, Charlie Frogner and all the other members of the Geometric Data Processing group at CSAIL for entertaining my crazy ideas and letting me distract them by talking about hadron jets for the past two years.

- Jesse Thaler, Patrick Komiske, and Eric Metodiev for engaging conversations about the physics behind the machine learning in this project. I wouldn't have discovered my passion for physics without them, and their support is greatly appreciated.

- Christie Hong, who has been a rock in the hard times. You always show me the bright side of everything, and I wouldn't be where I am without you.

Last, and most importantly, a huge thank you to my parents, who have believed in me since the beginning, and to whom I owe everything. I love you guys.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> "The laws of nature are constructed in such a way as to make the universe as interesting as possible."
>
> — Freeman Dyson, *Imagined Worlds*

1.1 Motivation

The search for physics beyond the Standard Model relies on particle accelerators to give us a window into a world of high energies, small distances, and short timescales. The effective theories governing this regime are strange and foreign in comparison with how we interact with matter in our terrestrial lives, but they provide a valuable insight into the fundamental question physicists have been asking since the dawn of time: *What are the equations that govern the universe?*

Currently, our best guess at a unified theory of everything is the Standard Model. But what lies beyond it? One way to probe new physics is by smashing together protons at insanely large energies to create new subatomic particles. In particular, when a proton undergoes scattering at energy scales on the order of tera-electron-volts, it will fragment into its constituent partons – three quarks and three gluons. Quarks and gluons are, under most circumstances, confined by the laws of quantum chromodynamics to be bound together into hadrons. Due to the asymptotic freedom of the strong force (meaning, roughly, that it decays at short length scales and high energies) these elementary building blocks can deconfine. In this scattering process,

they couple to each other and the vacuum to form new resonant particles. It is these anomalies that allow us to experimentally test new theories of the universe.

The pathway from proton collision to announcing a new particle has been discovered is not as easy as it may sound. As the scattering process happens at relativistic speeds, we can only observe the energy distribution deposited in the calorimeter around the cylindrical outside of the collision chamber. These event signatures are also messy and stochastic. Anecdotally, the process of proton collisions has been described as two people standing on opposite sides of a football field and throwing bowls of pea soup at each other at relativistic speeds. While one is interested in what happens when the peas collide, most of the time one just gets soup everywhere.

In general, it is very difficult to deterministically reconstruct the intermediate particles that generate a given event signature. However, we can gain valuable insights, potentially both discovering new particles and improving precision measurements of observable quantities, by analyzing the wealth of data provided by particle colliders like the LHC. To achieve the best understanding, it is crucial to leverage as much structure from the data as possible.

In this thesis, we will explore the space of anomaly detection techniques, with a primary focus on the geometric and statistical advantages conveyed by the structure of event signatures. As event signatures are distributions of energy over space, it is natural to ask under what conditions two distributions are similar or different. To answer that question, we will appeal to the theory of optimal transportation. Optimal transport provides a distance metric on the space of probability distributions defined over arbitrary measurable spaces. This framework allows us to utilize many tools in geometric machine learning to understand anomalous events, and motivates this thesis.

1.2 Notation

Coordinates. At the Large Hadron Collider, final state particles are traditionally recorded in one of two coordinate systems. Let the z-axis be longitudinal (i.e., along

the beam). The first coordinate system is $\vec{\mathbf{p}} = (p_x, p_y, p_z)$. Alternately, we sometimes write (p_T, η, ϕ):

$$\begin{aligned} p_T &= \sqrt{p_x^2 + p_y^2} \\ \eta &= \frac{1}{2} \ln \left(\frac{\|\vec{\mathbf{p}}\| + p_z}{\|\vec{\mathbf{p}}\| - p_z} \right) \\ \phi &= \sin^{-1} \left(\frac{p_y}{p_x} \right) \end{aligned}$$

The second set of coordinates has the benefit of transforming additively to Lorentz boosts along the longitudinal axis. For example, a boost of magnitude $\Delta\eta$ moves a massless particle to coordinates $(p_T, \eta + \Delta\eta, \phi)$. We will commonly use the Euclidean distance metric in $\eta - \phi$ space, defined as: $(\Delta R)^2 = (\Delta\eta)^2 + (\Delta\phi)^2$.

Measures and spaces. We will denote random variables by uppercase letters X. A random variable that follows a distribution law is written as $X \sim \mu$. For a metric space $\mathcal{X}$, we define the space of all continuous and infinitely differentiable functions $\mathcal{C}^\infty(\mathcal{X})$ and the set of Radon probability measures $\mathcal{P}_+(\mathcal{X})$. In the continuous case, the expectation of a function f of a random variable $X \sim \mu$ will be written $\mathbb{E}_\mu[f(X)] \triangleq \int_{\mathcal{X}} f(X) d\mu$. The variance will be written as $\mathsf{Var}[f(X)] = \mathbb{E}[f(X)^2] - (\mathbb{E}[f(X)])^2$. Given a map $f : \mathcal{X} \to \mathcal{Y}$, the push-forward operator is $f_\# : \mathcal{P}_+(\mathcal{X}) \to \mathcal{P}_+(\mathcal{Y})$. It sends the distribution μ defined over $\mathcal{X}$ to a distribution ν defined over $\mathcal{Y}$ where $(f_\# \circ \mu)(y) = \mu(f^{-1}(y))$.

When we transition to the discrete setting, the probability simplex on n bins will be written as $\Delta^n = \{\mathbf{a} \in \mathbb{R}_+^n | \sum_i \mathbf{a}_i = \mathbf{1}\}$. The empirical measure corresponding to sampling n random variables $\{x_1, ..., x_n\} \sim \mu$ will be written as $\hat{\mu}_n = \sum_i \delta_{x_i}$, where δ_x is the Dirac delta at location x.

Finally, for computational purposes, we will write discrete distributions as vectors. Vectors $\mathbf{v}$ and matrices $\mathbf{A}$ will be displayed in boldface. Denote the special vector of all ones $\mathbb{1}_n = (1, 1, ..., 1)^T \in \mathbb{R}^n$. The inner product is the standard Euclidean inner product for vectors $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}^T \mathbf{b} = \sum_i \mathbf{a}_i \mathbf{b}_i$, and similarly the Frobenius inner product for matrices $\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{ij} \mathbf{A}_{ij} \mathbf{B}_{ij}$.

1.3 Overview

The remainder of this thesis is organized into three different parts.

In Part I, we give brief preliminaries on particle physics and machine learning necessary to understand the remainder of this thesis.

- Chapter 2 delves into the experimental apparatus used to generate data at particle accelerators, and some theoretical formalisms necessary to understand it. We will also give a brief introduction to subatomic particles and jet physics, focusing on the perspective of data analysis.

- Chapter 3 introduces the challenge of learning on point clouds and provides some theoretical motivation for considering transport-based metrics. We also formally write the optimal transport problem, as well as describing the computational challenges inherent in solving it.

In Part II, we provide some theoretical improvements in efficiently computing Wasserstein-type kernels, which will be the building blocks underlying some of our anomaly detection techniques.

- Chapter 4 builds a connection between kernel approximations and geodesic clustering in a high-dimensional feature space. We show a quadrature bound for coresets of arbitrary functions over a measure that is inspired by the idea of Wasserstein barycenters and centroidal Voronoi tessellations.

- Chapter 5 covers two techniques for more efficiently computing transport-type distances. We describe an unbiased version of the sliced Wasserstein distance, and give two formulations for computing it that reduce variance. We also demonstrate a multi-level Monte Carlo approximation and prove that our estimator enjoys better scaling than existing techniques. Finally, we introduce kernels based on these metrics.

In Part III, we apply our techniques to perform anomaly detection on simulated collider data. We will outline and test two types of methods: discriminative methods,

which classify individual datapoints as anomalous or not, and generative methods, which seek to statistically model the properties of anomalies in the aggregate.

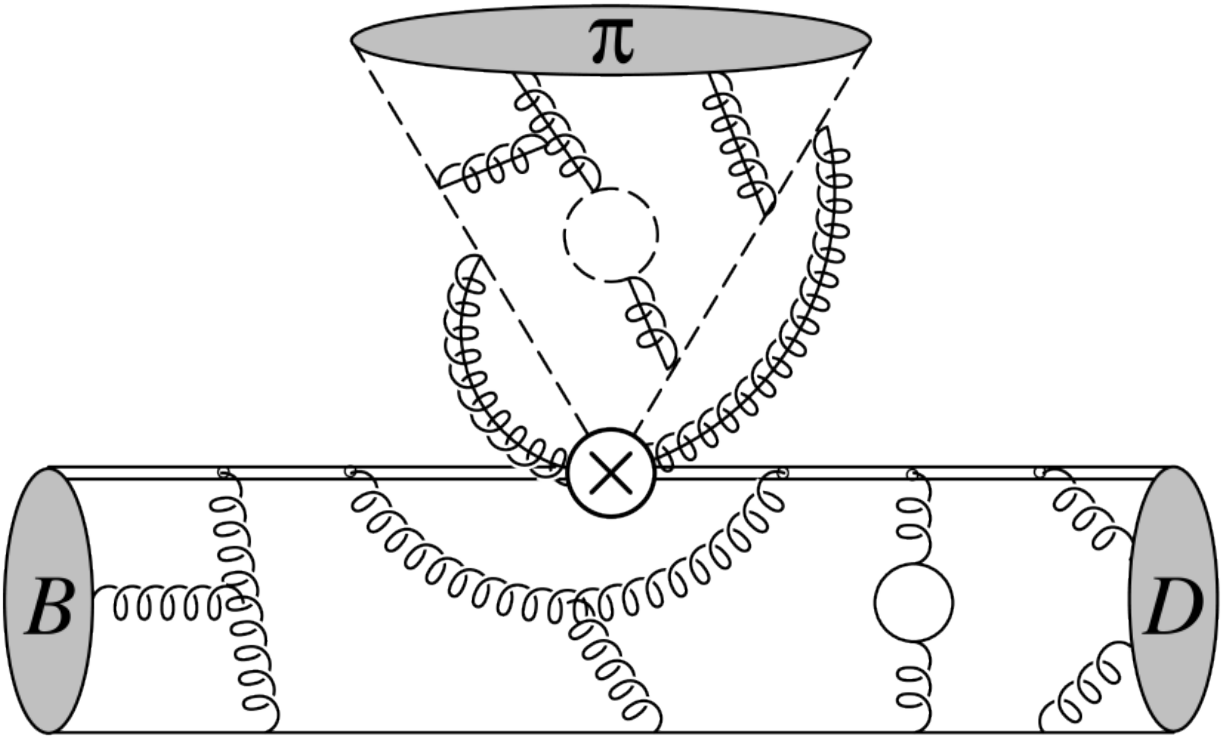
- Chapter 6 describes the application of two discriminative methods to anomaly detection problems. We perform an empirical study of the various types of kernel estimators developed in the previous chapter. Additionally, we compare density-based and margin-based kernelized anomaly detection techniques on a top quark tagging dataset.

- Chapter 7 applies a domain-specific generative technique based on topic modeling to simulated LHC data. We show good performance of our method to find resonant anomalies, as well as to discriminate between quark and gluon-initiated jets.

Chapter 8 provides concluding remarks.

Part I

Background



The intersection of machine learning and particle physics is an exciting and rapidly-developing field. The wealth and richness of data collected from the Large Hadron Collider allows us to peer deeper into the building blocks of the universe and test more and more complex theories of physics beyond the Standard Model. In this thesis, we will primarily focus on *jet physics*, one of the oldest aspects of this vast field. Briefly, jets are collimated sprays of subatomic particles generated when protons smash into each other at relativistic speeds.

While there is no theoretical definition of what constitutes a jet, in practice, they are some of the most fundamental objects of study in collider data. As an example, understanding the scattering processes that lead to the creation of a jet is an active field of study. The figure on the previous page, taken from ref. [6], provides an example of a Feynman diagram for a soft collinear coupling that affects the intermediate states. Similar processes cause the Casimir scaling effect between quark and gluon jets, to which we will return in the last chapter of this thesis.

This part serves as a brief introduction to the methods of the rest of this thesis. Chapter 2 will introduce the Standard Model, with the goal of explaining how data is generated and collected at the Large Hadron Collider. Chapter 3 is devoted to understanding the computational and mathematical challenges of modelling collections of unordered point clouds, with a focus on optimal transport and related techniques.

Chapter 2

Accelerator physics

"The Standard Model is so complex it would be hard to put it on a T-shirt – though not impossible; you'd just have to write kind of small."
— Steven Weinberg

2.1 The Standard Model

The Standard Model is the crowning achievement of almost a century of progress in understanding the fundamental laws governing our universe [7]. It can be succinctly summarized by the following Lagrangian [8]:

$$\mathcal{L} = -\frac{1}{4}(F^a_{\mu\nu})^2 + \bar{\psi}(i\gamma^\mu D_\mu)\psi + y_{ij}\bar{\psi}_i\psi_j\phi + |(\partial_\mu - igA^a_\mu t^a_r)\phi|^2 + \mu^2\phi^\dagger\phi - \lambda(\phi^\dagger\phi)^2$$

It describes three of the four fundamental forces in our universe: electromagnetism, and the strong and weak nuclear forces.[1] Of these three forces, electromagnetism is the only one that is noticeable on a terrestrial length scale.

To understand how the SM relates matter and force, we must retrace the steps of Max Planck in discovering the quantization of radiation. Briefly put, Planck accounted for the existence of blackbodies by proposing that electromagnetic radiation

[1] Notably, gravity is not included in the Standard Model. Finding a theory to unify gravity with the rest of the SM is one of the most pressing open questions in theoretical physics. For more information, we refer the reader to ref. [9]. For the purposes of this thesis, gravity is negligible at the energy scales of particle colliders.

could only be emitted in discrete *quanta*. A quantum is defined as the smallest discrete amount of any property that is permitted to participate in an interaction. The electromagnetic force is propagated by a massless particle called a photon, which is denoted γ. The position and momentum of any photon cannot be determined precisely. Instead, these observables are governed by a probability distribution. Further, the relative precision to which we can measure canonically conjugate pairs of observables is constrained.[2] This simple statement demonstrates that the underlying degrees of freedom dictating observable events are non-deterministic. Hence, when we refer to particles in this thesis, we do not mean point masses in the classical sense. We instead speak of particles as the propagators of *quantum fields*, which are operator functions defined at every point on spacetime.[3]

2.1.1 An inventory of particles

Each fundamental force in the Standard Model is associated with both a quantum field and a mediator particle which governs its interactions. A full inventory of these fields and particles is given in Table 2.1. For example, the photoelectric effect is the interaction of photons in the form of incoming radiation with electrons bound to nuclei in a metal, and represents one mode of interaction of the electromagnetic force. Similarly, the strong and weak nuclear forces are also mediated by subatomic particles. For the weak nuclear force, these are the $\mathsf{W}^+, \mathsf{W}^-$, and Z bosons; for the strong force, they are the 8 different types of gluons g. Collectively, these particles are known as the *gauge bosons*. Of these bosons, only the weak force mediators have non-zero mass (the W^+ has mass roughly 80.3 GeV). For this reason, the weak force decays very rapidly with distance – each boson has a half-life of less than 3×10^{-25} seconds [11]. Finally, we must mention possibly the most famous boson – the Higgs. The Higgs boson is a scalar boson, meaning it has zero spin, unlike the others, which have spin 1. The Higgs field is special due to its non-zero *vacuum expectation value* –

[2] A more familiar version of this statement is given by the Heisenberg uncertainty principle.

[3] The full complexity of quantum field theory is too great to describe in this brief introduction. There are an innumerable number of textbooks on QFT, but we recommend [10] for a intuitive, if not rigorous introduction.

that is, the average value of the Higgs operator in the absence of other particles. This property leads to electroweak symmetry breaking, allowing particles to gain potential energy through coupling. This is what we commonly refer to as mass. The Higgs field is particularly important as it represents the greatest triumph of the Large Hadron Collider, as well as the last major piece of the Standard Model to be experimentally verified. We will return to the Higgs field in section 2.2, when we describe how it was experimentally verified in 2012 [4].

Field	Particles	Representation	Generations
Spin 1 gauge bosons			
B	photon (γ)	$(1, 1, 0)$	
W	W^+, W^-, Z	$(1, 3, 0)$	
G	8 gluons (g)	$(8, 1, 0)$	
Spin $\frac{1}{2}$ fermions			
q_L	quarks (u,d,c,s,t,b)	$(3, 2, \frac{1}{3})$	3 total
$\bar{u}_L^C$	up antiquarks ($\bar{u}, \bar{c}, \bar{t}$)	$(\bar{3}, 1, -\frac{4}{3})$	3 total
$\bar{u}_L^C$	down antiquarks ($\bar{d}, \bar{s}, \bar{b}$)	$(\bar{3}, 1, \frac{2}{3})$	3 total
$\bar{\ell}_L$	leptons (e^-, μ^-, τ^-)	$(1, 2, -1)$	3 total
$\bar{\ell}_L^C$	antileptons (e^+, μ^+, τ^+)	$(1, 1, 2)$	3 total
Spin 0 scalar bosons			
H	Higgs boson (H)	$(1, 2, 1)$	

Table 2.1: The particle inventory of the Standard Model. The representation column gives the set of transformations in $SU(3) \times SU(2) \times U(1)$ under which each particle transforms.

The Standard Model also specifies a different class of particle, known as the *fermions*. As we will see, matter itself is an emergent phenomenon, composed of the interactions of this second class of particles mediated by the bosons. Fermions carry *charge*, and act as source generators for the field associated with a charge. Charge is the generator of a symmetry group acting on a field. For example, the electron is a member of the subclass of fermions known as *leptons* (from the Greek "leptos" for their light mass). The electron carries a negative electric charge, and interacts with other particles via the electromagnetic force. The symmetry group acting on the EM

field is $\mathsf{U}(1)$, meaning that there is only one type of electric charge. By comparison, the other subclass of fermions are the *quarks*. Quarks carry both electric charge and the strong-force equivalent of charge, known as *color*. There are three types of color charge, known as red, blue and green,[4] each corresponding to a generator of the $\mathsf{SU}(3)$ symmetry of the SM. For completeness, we note that the weak-force charge is known as *weak isospin*, corresponding to $\mathsf{SU}(2)$ symmetry. All massive particles carry weak isospin and, therefore, interact with the W and Z bosons. Quarks are the constituent components of protons and neutrons. In total, there are 6 quarks, organized into three "generations": up (u) and down (d), charm (c) and strange (s), top (t) and bottom (b). and their corresponding antiparticles. Both protons and neutrons are *hadrons*, composite particles made of quarks bound by gluons. A proton consists of two up and one down quark (uud), while a neutron is (udd). Protons exist in a color singlet state, given by a color charge of $(\mathsf{r\bar{r}} + \mathsf{b\bar{b}} + \mathsf{g\bar{g}})/\sqrt{3}$. This means that in any proton, each color must be represented by exactly one quark [12].

2.2 The Large Hadron Collider

The Large Hadron Collider is the world's largest particle accelerator. It consists of over 30 kilometers of track, and has been operating for the past 12 years under Geneva. During operation, the superconducting magnets around the collider accelerate two beams of protons to super-relativistic speeds. Each beam consists of approximately 1,000 bunches, each containing 100 billion protons. Head-on collisions between bunches occur every 25 nanoseconds, at center-of-mass energies of $\sqrt{s} = 13$ TeV [8]. To put this into perspective, this means each bunch of protons has roughly the same kinetic energy as a small anti-tank explosive.

For our purposes, the most important interactions are those that occur within a proton. Quarks and gluons collectively are known as *partons*, a term due to Feynman. Partons, and, more generally, any particles carrying color charge experience a peculiar

[4] These have nothing to do with our common perception of color. Physicists are just bad at naming things.

phenomenon known as *color confinement*. In simple terms, confinement states that only colorless particles – those that have no net color charge – are stable. Most of the mass of a proton (i.e., its rest energy) comes from the chromodynamic binding energy of the gluons which mediate the strong force between the quarks.

Confinement is broken at extremely high energy scales due to the asymptotic freedom of the strong interaction [13]. When gluons bind with each other in vacuum, they cause a quasi-paramagnetic polarization in color, which in turn makes the vacuum a quasi-dielectric in color. Thus, unlike in electromagnetism, the vacuum is "anti-screening" to the strong force, and its interactions weaken at extremely small distances. Hence, in a high-energy collision, as the quarks and gluons move close to each other, they uncouple and become asymptotically free to move in space. This process is known as *fragmentation*. The bare quarks and gluons are exposed for a fraction of a second, but due to confinement, these isolated particles are not stable. Hence, they must *hadronize* into composite particles. It must be noted that a solid theoretical understanding of these processes does not exist; while the mechanism that causes confinement has been extensively studied [14, 15], there is no known non-abelian gauge theory that is guaranteed to have confinement. Models that predict hadronization well in practice, however, do exist [16].

After fragmentation, hadrons may not necessarily be stable particles. In particular, the final state might go through an arbitrary sequence of intermediate products. However, the intermediate state is never directly observed, as the decay happens much too quickly to detect. Instead, colorless final-state hadrons are recorded in a series of multiple types of calorimeters surrounding the detector, as shown in Fig. 2-1.

Equipped with this understand, we can now define an event, the central object of study for accelerator physics.

Definition 2.2.1. (*Events at the LHC.*) An event $\mathcal{E}$ is an unordered collection of reconstructed particles $\{(p_T, \eta, \phi)_i\}_{i=1}^{N}$, derived from calorimeter readings. It is an empirical approximation to an underlying energy distriution.

For our purposes, all particles in the reconstruction are both massless and colorless.

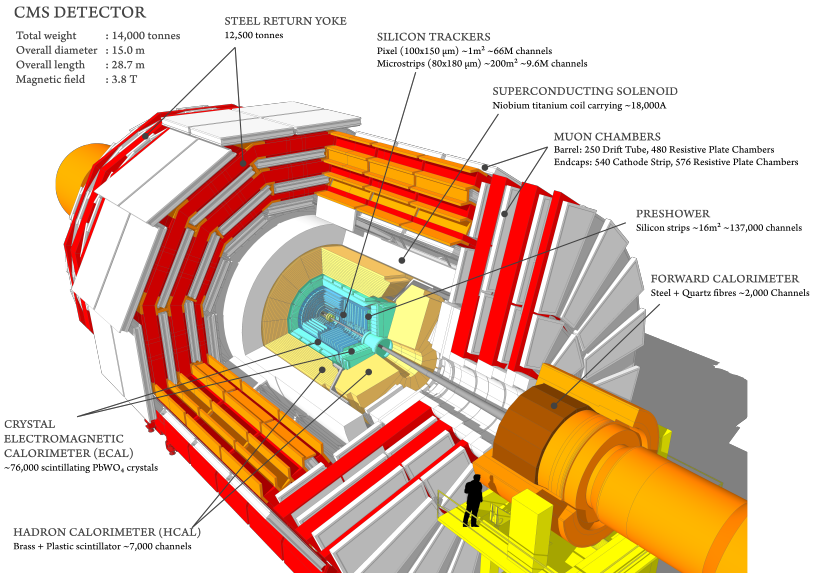


Figure 2-1: A cutaway graphic showing the structure of the CMS (Compact Muon Solenoid) detector at the Large Hadron Collider. Figure taken from the CMS Collaboration website.

This means that the event signature contains no categorical variables – each particle in the event signature is uniquely identified by a spatial location and an energy or momentum. In particular, this suggests that events can be considered as point clouds. While we will discuss point clouds in more detail in Chapter 3, we note that they can alternately be thought of as empirical approximations generated by sampling from some underlying continuous distribution; this point of view (events as energy distributions) will be one we adopt for the remainder of this thesis.

2.3 Jet kinematics

Precision measurements of SM parameters and probes of beyond-SM physics both require accurate understanding of the intermediate channels by which final-state hadrons are produced. Conveniently, most intermediate products have large transverse momentum as they decay and hadronize. Further, the likelihood for creating a new particle decreases with increasing scattering angle. As a result, the bremsstrahlung products they form as they decelerate through the vacuum (gluons, quark-antiquark pairs, etc.) are highly collimated. Hence, the structure in $\eta - \phi$

space of the final event is well-correlated with the intermediate structure. These highly collimated regions of phase space are known as jets.

Definition 2.3.1. (*Jets.*) A jet $\mathcal{J}$ is a subset of an event $\mathcal{E}$ corresponding to a localized and highly collimated spray of particles. There is no universal definition of a jet; however, operationally, a jet is a reconstructed region of the event corresponding to a discrete energy flow.

Jets are a fundamental object of study in both precision measurements and searches of beyond-Standard Model physics at the Large Hadron collider. Many techniques exist for reconstructing jets. A full review is given in ref. [17]. For the purposes of this thesis, we will only consider the anti-k_t algorithm due to ref. [1], which is a special case of the sequential recombination algorithms. An example of the application of this algorithm is given in Fig. 2-2.

Given a radius parameter R, define the metric between particles and the distance to the beam axis as follows:

$$d(\vec{p}_i, \vec{p}_j) \triangleq \min(p_{T,i}^{-2}, p_{T,j}^{-2}) \frac{(\Delta R)^2}{R^2} \tag{2.1}$$

$$d(\vec{p}_i, \mathcal{B}) \triangleq p_{T,i}^{-2} \tag{2.2}$$

The algorithm proceeds as follows:

1. For each particle $x_i \in \mathcal{E}$, create a new proto-jet and assign the particle to it.

2. While there are still proto-jets, find the pair of proto-jets (k, ℓ) with the minimum distance between them $d(\vec{p}_k, \vec{p}_\ell)$, and the single proto-jet m with the minimum distance to the beam axis $d(\vec{p}_m, \mathcal{B})$. If $d(\vec{p}_m, \mathcal{B}) < d(\vec{p}_k, \vec{p}_\ell)$, define m to be a jet and remove it from the set of proto-jets; else, define a new proto-jet with momentum $\vec{p_n} = \vec{p_k} + \vec{p_\ell}$.

Given a single jet, many observables can be used to understand its properties. We will briefly outline a few relevant ones here, but leave an exhaustive review of jet observables to ref. [18].

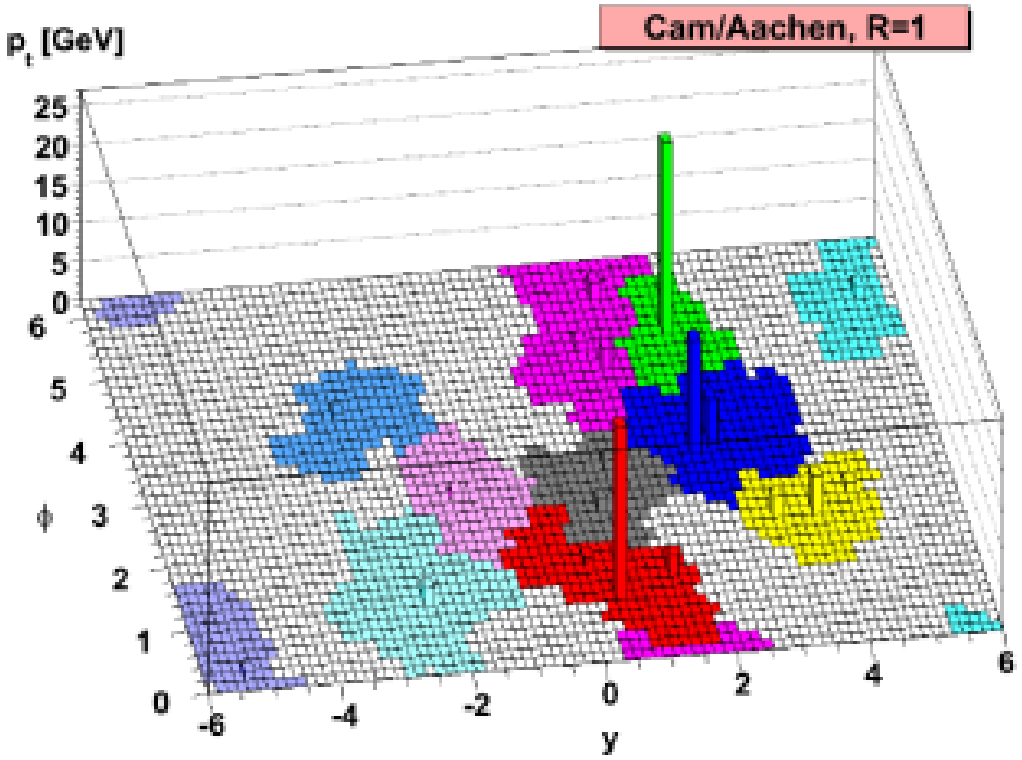


Figure 2-2: An example application of the anti-k_t algorithm, with characteristic radius 1. Note the conical form of the jets produced. Figure from ref. [1].

1. **Jet mass.** The invariant jet mass for a jet $\mathcal{J}$ is defined as the Minkowski norm of the sum of the 4-momenta of all the particles in the jet. Concretely, it is calculated as

$$m(\mathcal{J}) = \sqrt{\left(\sum_{i \in \mathcal{J}} E_i\right)^2 - \left\|\sum_{i \in \mathcal{J}} \vec{p_i}\right\|^2}$$

where E is the energy recorded in the calorimeter.

2. **n-subjettiness.** First introduced in [19], this correlates with the number of "prongs" or modes in the energy distribution of a jet. Lorentz-boosted bosons often are 2-subjetty, while boosted top quarks are 3-subjetty.

3. **Jet multiplicity.** Defined as the number of total particles detected in a jet. As outlined in [20], jets that are quark-initiated have different multiplicity distributions than those that are gluon-initiated. We will return to this distinction in Chapter 7.

Many other observables have been proposed in the literature for understanding specific types of jets with and without domain-specific knowledge [21, 22]; however, for the remainder of this thesis, we will focus primarily on jet mass and multiplicity.

2.4 Searches at the LHC

Studying reconstructed jets gives us insight into the intermediate channels of production of rare particles at the LHC. In particular, recent developments in jet substructure both in theory [23–30] and in experiment [31–39] have proved extremely promising for a variety of different applications in accelerator physics. We will take as an exemplar search the successful discovery in 2012 of the Higgs boson at the LHC [4]. As summarized in Figure 2-3, there are multiple channels by which the Higgs boson can be created in pp collisions. As proton quarks (the u, d) form the lightest generation, and the Higgs coupling increases with mass, the most common channel to create the Higgs is in fact the gluon fusion channel denoted as (a) in the figure below.

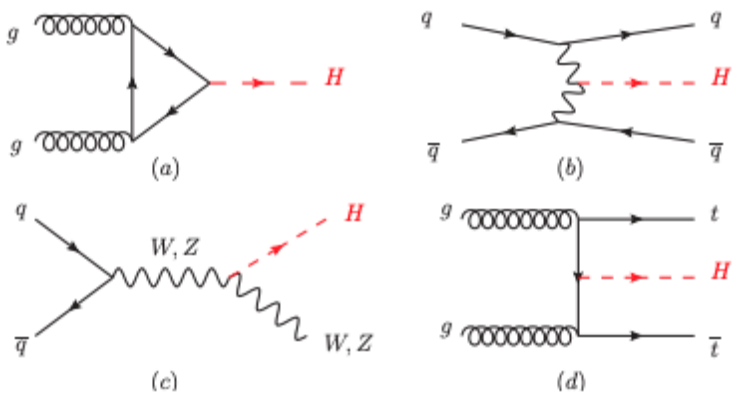


Figure 2-3: The Feynman diagrams corresponding to the main channels of production of the Higgs boson at the LHC. (a) gluon fusion. (b) weak boson fusion, (c) production via gauge boson, (d) production via top quark. Figure from ref. [2].

Yet, more important for the purposes of our study are the *decay channels* for the Higgs. There are three dominant modes by which the Higgs decays: $\mathsf{H} \to \mathsf{VV}$, $\mathsf{H} \to \mathsf{gg}$, $\mathsf{H} \to \mathsf{q\bar{q}}$, where $\mathsf{V}, \mathsf{g}, \mathsf{q}$ are weak gauge bosons, gluons, and quarks, respectively. This is evidenced in Fig. 2-4.

Most proton collisions result in two individual and well-separated jets. When the density any observable of these generic QCD jets is plotted, the histogram forms a smoothly falling background, as shown in the top panel of Fig. 2-5. However, because the Higgs boson has a well-defined mass, the jets created by the Higgs maintain the same invariant mass, with some added noise. This results in a Breit-Wigner distribution centered around the true Higgs mass. In the histogram, it appears as a small bump. The presence of this bump at a statistically significant level is evidence for the presence of the Higgs. The notion of anomalies as "relative overdensities" in a region of some well-defined phase space is central to search for new physics, as well as the rest of this thesis.

2.5 Machine learning in BSM physics

In searches for physics beyond the SM, the resonant mass of a certain particle may not be known *a priori*. Our work in this thesis attacks two important tasks crucial

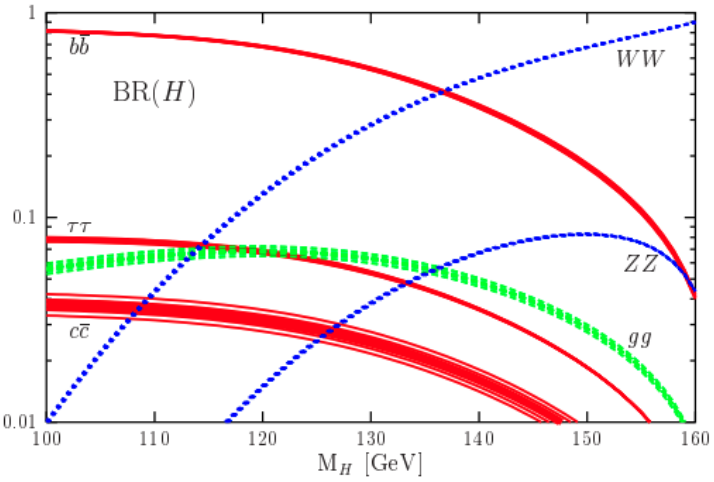


Figure 2-4: Branching ratios for Higgs decay channels at the Large Hadron Collider. At the currently accepted mass $m_H = 125.18 \pm 0.16$ GeV, the dominant decay modes are 2-pronged jets. Figure from ref. [3].

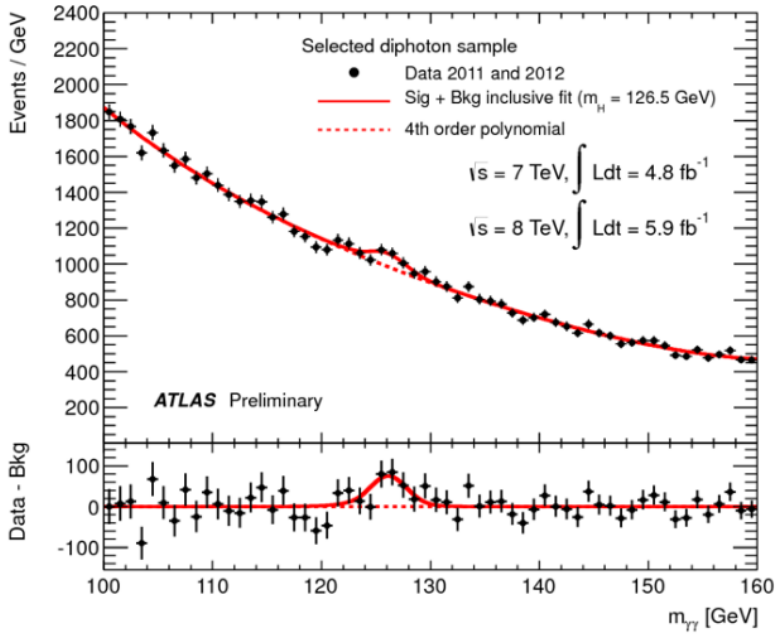


Figure 2-5: The bump in the invariant mass spectrum corresponding to the anomalous Higgs excess along the diphoton decay channel. Figure from ref. [4].

to BSM probes.

First, the task of *resonant anomaly detection* is to find, without assumptions on the physical characteristics of a resonance, distributional excesses corresponding to unknown particles. Many techniques for detecting these anomalies have been pro-

posed. Some are based on sophisticated deep learning techniques like the variational autoencoder, a method which seeks to learn a low-dimensional representation of data; the anomalies are the events which have a high reconstruction error [40, 41]. Other techniques seek to improve the statistical sensitivity of traditional "bump hunting" methods [42, 43]. Our approach will rely on the mathematical model of an event as a distribution of energy across space. To perform anomaly detection, we will leverage techniques from *optimal transport* theory [44, 45], first applied to this field in ref. [46]. In the next chapter, we will introduce optimal transport and outline why it is useful for defining a metric on the space of collider events.

Second, we will build a model to discriminate between quark- and gluon-initiated jets. By classifying light jets in this way, we are able to gain insight into the dominant decay channels for an unknown intermediate resonance. Many machine learning classifiers have been proposed for this problem, ranging from deep neural networks [47–50] to cutting distributions based on simple and complex observables [51, 52]. Our work falls into the class of *generative models*, first applied to quark-gluon discrimination by ref. [53]. Our contribution is a generalization of their model to dijet resonances by leveraging the factorization theorem for dijets.

Chapter 3

Notions of distance

"The first good thing about optimal couplings is that they exist."

— Cédric Villani, *Optimal transport, old and new*

In this section, we will address the challenge of defining when two distributions are similar and when they are not. While this may seem simple at first glance, this problem reveals a grand mathematical formalism that is in equal parts subtle and powerful. This section will focus primarily on the applied side of the field, when we only have access to probability distributions through sampling. We call these *point clouds*, and we will study them from the statistical perspective where they are viewed as empirical approximations to some underlying measure. First, we will introduce the concept of a φ-divergence, a class that encompasses many commonly used notions of distance between distributions employed in the statistical and machine learning literature. As we will see, this class of divergences lacks certain properties that would be desirable when it comes to the discrete and empirical settings. For a class of distances avoiding these issues, we will turn to optimal transport and the Wasserstein distances. This framework will be the building block for the kernels we will use in our anomaly detection techniques later in this thesis.

3.1 Discrepancies between measures.

3.1.1 φ-divergences

Before we move to the discrete setting, we first mention several notions of discrepancy between absolutely continuous measures. These are the φ-divergences first introduced by [54], defined as follows:

Definition 3.1.1. (φ-*divergence.*) Let φ be a convex and continuous function such that $\varphi(1) = 0$. The φ-divergence is between two measures $\mu, \nu \in \mathcal{M}_+^1(\mathcal{X})$ as:

$$D_\varphi(\mu \| \nu) \equiv \int_{\mathcal{X}} \varphi\left(\frac{d\mu}{d\nu}\right) d\nu$$

In particular, consider the ℓ_1 (or total-variation) distance and the Kullback-Leibler (KL) divergence, both of which are commonly used to define similarity of probability distributions. They are defined in Table 3.1. As shown, both are φ-divergences. However, both these divergences are ill-behaved when the supports of the distributions being compared are different. In particular, if $\mathsf{supp}(\mu) \cup \mathsf{supp}(\nu) \subset \mathcal{X}$ is a set of measure zero, then $\mathsf{KL}(\mu \| \nu) = \infty$ and $\mathsf{TV}(\mu, \nu) = C$ for some positive constant C. This statement holds regardless of how similar or different μ, ν are, both in distributional shape and location of support in the ambient space $\mathcal{X}$. As an example, consider the scenario where $\mathcal{X} = \mathbb{R}$, $\mu = \delta_0$ and $\nu = \delta_x$ for some $x > 0$. It is problematic that the value of x does not affect the value of the φ-divergence.

Name	Divergence	$\varphi(\cdot)$
Total variation	$\mathsf{TV}(\mu, \nu) = \sup_{\mathcal{A} \subseteq \mathcal{X}} \|\mu(\mathcal{A}) - \nu(\mathcal{A})\|_1$	$\varphi_{\mathsf{TV}}(t) = \frac{1}{2}\|t - 1\|$
Kullback-Leibler (KL)	$\mathsf{KL}(\mu \| \nu) = \int_{\mathcal{X}} \log \frac{d\mu}{d\nu} d\mu$	$\varphi_{\mathsf{KL}}(t) = t \log t$

Table 3.1: Two common φ-divergences, and their respective functions $\varphi(\cdot)$.

For the purposes of this thesis, we can consider a point cloud $\hat{\mu}_n$ to be an empirical approximation to an absolutely continuous distribution μ. However, the argument

above suggests that $D_\varphi(\hat{\mu}_n, \hat{\nu}_n)$ will be degenerate, even if $\mu = \nu$. To make this statement more formal, we state the following result from ref. [55].

Definition 3.1.2. *(Weak convergence.)* A sequence of measures $\{\mu_n\}_{n=1}^{\infty}$ converges weakly to a reference measure μ (written as $\mu_n \xrightarrow{\mathcal{D}} \mu$) if the following statement is true:

$$\lim_{n\to\infty} \sup \mathbb{E}_{\mu_n}[f(X)] = \mathbb{E}_{\mu}[f(X)]$$

for every continuous and bounded function $f : \mathcal{X} \to \mathbb{R}_{\leqslant c}$. A divergence metrizes weak convergence if:

$$D(\mu_n, \mu) \to 0 \iff \mu_n \xrightarrow{\mathcal{D}} \mu$$

Thus, the φ-divergences do not metrize weak convergence. We turn instead to a different notion of distance based on a principle of least action.

3.2 Optimal transport

Consider two distributions μ, ν that are represented as piles of sand. Moving sand from one spatial location to another costs energy. What is the minimum amount of energy required to fully transform μ into ν? The answer to this question is the *optimal transport plan*. Transport has found applications in a variety of disparate fields, from economics [56, 57] to computer vision [58–61] to astrophysics [62, 63]. In this section, we will give a brief introduction to the continuous and discrete versions of the problem.

3.2.1 Exact formulation

The exact optimal transport problem can be formulated as follows [45], first due to Leonid Kantorovich:

Definition 3.2.1. *(Optimal transportation.)* Given two measures μ, ν on metric spaces $\mathcal{X}, \mathcal{Y}$ respectively, and a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geqslant 0}$, the smallest transportation cost required to move μ to ν is given by:

$$\mathcal{W}_c(\mu,\nu) \equiv \inf_{\gamma\in\Gamma(\mu,\nu)} \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\,d\gamma(x,y) \right\}$$

where $\Gamma(\mu,\nu)$ is the set of couplings, i.e., probability distributions on $\mathcal{X}\times\mathcal{Y}$ such that the marginals of $\gamma\in\Gamma$ on $\mathcal{X},\mathcal{Y}$ are μ,ν respectively.

As alluded to in the epigraph, it can be shown that the minimizing transport plan always exists [55]. To show this, we transform to the dual of the linear program above. Note the corresponding dual problem can be written as a supremum over expectations of the *Kantorovich potentials*:

$$\mathcal{W}_c(\mu,\nu) = \sup_{f,g\in\Phi} \int_{\mathcal{X}} f\,d\mu + \int_{\mathcal{Y}} g\,d\nu$$

where $\Phi = \{(f,g) \in \mathcal{C}^\infty(\mathcal{X}) \times \mathcal{C}^\infty(\mathcal{Y}) : \forall x,y, f(x) + g(y) \leqslant c(x,y)\}$ is the set of admissible dual potentials. Obtaining the transport map is then a matter of taking first variations of the dual potentials.[1]

When the cost $c(x,y)$ equals $\|x-y\|_p^{1/p}$, the optimal transport cost is known as the p-Wasserstein or $\mathcal{W}_p$ distance. In the remainder of this work, we use the Wasserstein distance $\mathcal{W}_2$, and the metric spaces we consider are Euclidean ($\mathcal{X}=\mathcal{Y}=\mathbb{R}^d$). Finally, we consider μ,ν to be discrete distributions and treat them as vectors $\boldsymbol{\mu} = \{\mu_i\}_{i=1}^M$. Thus, the problem can be rewritten in primal and dual forms as:

$$\boxed{\begin{array}{lll} \min\limits_{\mathbf{T}} & \langle \mathbf{C}, \mathbf{T}\rangle & \\ \text{s.t.} & \mathbf{T}\mathbb{1}_n = \boldsymbol{\mu} & \\ & \mathbf{T}^\intercal \mathbb{1}_n = \boldsymbol{\nu} & \\ & \mathbf{T} \geqslant 0 & \end{array} \quad \Longleftrightarrow \quad \begin{array}{ll} \max\limits_{\mathbf{f},\mathbf{g}} & \langle \mathbf{f},\boldsymbol{\mu}\rangle + \langle \mathbf{g},\boldsymbol{\nu}\rangle \\ \text{s.t.} & \mathbf{f}\oplus\mathbf{g} \leqslant \mathbf{C} \end{array}} \tag{3.1}$$

where $\mathbf{C}$ is a matrix of pairwise distances, $\mathbf{T}$ is the transport plan, and $\mathbf{f},\mathbf{g}$ are the Kantorovich potentials. A visualization of this transport plan for a discrete distribution is given in Figure 3-1.

[1] There are many ways to show this result. For a full review of the subject, we refer the reader to ref. [45].

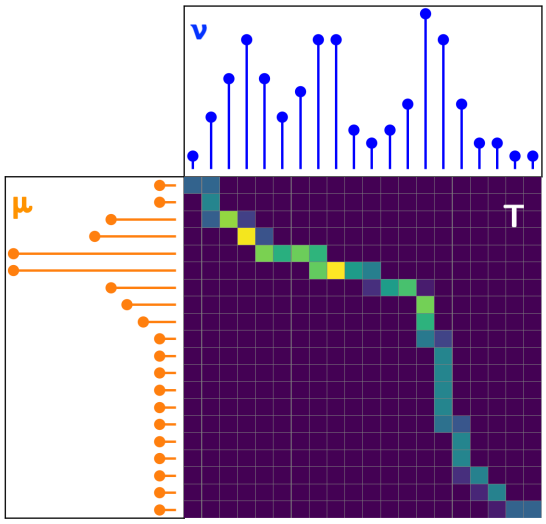


Figure 3-1: The optimal transport plan $\mathbf{T}$ under Euclidean cost between two discrete distributions. Figure adapted from ref. [5].

When $\mathcal{X} = \mathbb{R}$, the solution to the above linear program has a convenient closed form. Define the cumulative distribution function $\mathcal{Q}_\mu : \mathbb{R} \to [0, 1]$ such that $\mathcal{Q}_\mu(x) \triangleq \int_{-\infty}^{x} d\mu$. Similarly, define its pseudoinverse $\mathcal{Q}_\mu^{-1}$, the generalized quantile function. Then, the 1-dimensional transport distance is given by:

$$\mathcal{W}_2^2(\mu, \nu) = \int_0^1 \|\mathcal{Q}_\mu^{-1}(t) - \mathcal{Q}_\nu^{-1}(t)\|_2^2 dt \tag{3.2}$$

In the discrete setting, the above formula reduces to sorting. Assume the points are ordered such that $x_1 \leqslant x_2 \leqslant ... \leqslant x_n, x_i \sim \mu$ and similarly for $y_i \sim \nu$. Then, the following relation holds:

$$\mathcal{W}_2^2(\boldsymbol{\mu}, \boldsymbol{\nu}) = \frac{1}{n} \sum_{i=1}^{n} \|x_i - y_i\|_2^2 \tag{3.3}$$

However, in dimension $d \geqslant 2$, the simple closed form does not exist. Worse still, the exact solution of optimal transport suffers from two major problems:

- **Slow to compute.** This linear program has runtime complexity $\mathcal{O}(n^3 \log n)$ using traditional LP solvers [64]. Therefore, exact solutions to the OT linear program are prohibitive for large point clouds.

- **Slow to converge.** The convergence rate in terms of the number of samples is extremely slow, especially so for high-dimensional distributions. In particular, minimax results due to ref. [65] and others suggest that

$$\mathbb{E}\left[\left|\mathcal{W}_2^2(\hat{\mu}_n, \hat{\nu}_n) - \mathcal{W}_2^2(\mu, \nu)\right|\right] = \mathcal{O}(n^{-1/d})$$

  when $d > 4$. We will return to this more in Chapter 5.

For these reasons, we consider certain relaxations of the constraints to improve the computational properties of optimal transport.

3.2.2 Entropic optimal transport

Many techniques exist to improve the time complexity of computing optimal transport, the most popular by far of which is the Sinkhorn method [66]. This technique relies on adding an entropic penalty to the formulation of OT. Define the entropy of a matrix $\mathbf{P}$ to be $\mathbf{H}(\mathbf{P}) \triangleq -\sum_{i,j} \mathbf{P}_{ij}(\log \mathbf{P}_{ij} - 1)$. With a slight rewriting of Problem 3.1, we get the following:

$$\begin{aligned} \mathcal{W}_{\lambda,c}(\mu, \nu) = \min_{\mathbf{T}} \quad & \langle \mathbf{C}, \mathbf{P} \rangle - \frac{1}{\lambda}\mathbf{H}(\mathbf{T}) \\ \text{s.t.} \quad & \mathbf{T}\mathbb{1}_n = \boldsymbol{\mu} \\ & \mathbf{T}^\intercal \mathbb{1}_n = \boldsymbol{\nu} \\ & \mathbf{T} \geqslant 0 \end{aligned} \tag{3.4}$$

The solution to this equation has the form $\mathbf{T}_{ij} = \mathbf{u}_i \mathbf{K}_{ij} \mathbf{v}_j$ for some scaling vectors $\mathbf{u}, \mathbf{v}$. The celebrated Sinkhorn-Knopp matrix scaling algorithm allows for the following iterative update scheme [67]:

$$\begin{aligned} \mathbf{u}^{(\ell+1)} &\leftarrow \frac{\boldsymbol{\mu}}{\mathbf{K}\mathbf{v}^{(\ell)}} \\ \mathbf{v}^{(\ell)} &\leftarrow \frac{\boldsymbol{\nu}}{\mathbf{K}\mathbf{u}^{(\ell+1)}} \end{aligned}$$

As the regularization constant becomes large, the entropic Wasserstein distance converges to the exact solution of the linear program given in Problem 3.1:

$$\lim_{\lambda\to\infty} \mathcal{W}_\lambda(\mu,\nu) = \mathcal{W}(\mu,\nu)$$

Recent developments [5, 68, 69] in analyzing the Sinkhorn iterations for optimal transport have shown that the worst-case runtime for this algorithm is $\tilde{\mathcal{O}}(n^2\epsilon^{-2})$, where ϵ is the error relative to the exact cost.

3.3 Optimal transport in particle physics

Over the past several years, optimal transport and related techniques have demonstrated significant promise in analyzing jets and events in hadron colliders. In this section, we will briefly review some of the theoretical properties that make optimal transport a valuable tool for this field. While we will not engage with the definitions in this section in detail, we will note that all the estimators we propose in Chapter 5 also enjoy these properties. Starting with ref. [70], an unbalanced version of the 1-Wasserstein distance was shown to have certain beneficial properties with respect to the underlying physics of hadron collisions. In particular, this distance, which the authors refer to as the Energy Mover's distance [58] satisfies the important properties of infrared safety and collinearity (IRC).

Definition 3.3.1. (*IRC safety.*) A distance between events $\mathcal{E}_1, \mathcal{E}_2$ is IRC-safe if it meets the following two conditions:

1. **Infrared safety.** A distance is infrared-safe if:

$$\lim_{\epsilon\to 0} D\Big((1-\epsilon)\cdot\mathcal{E}_1 + \epsilon\cdot\delta_{x'}, \mathcal{E}_2\Big) = D(\mathcal{E}_1,\mathcal{E}_2)$$

   regardless of the location of emission x'. In words, adding a soft emission with infinitesimal energy should not change the distance.

2. **Collinear safety.** A distance is collinear-safe if:

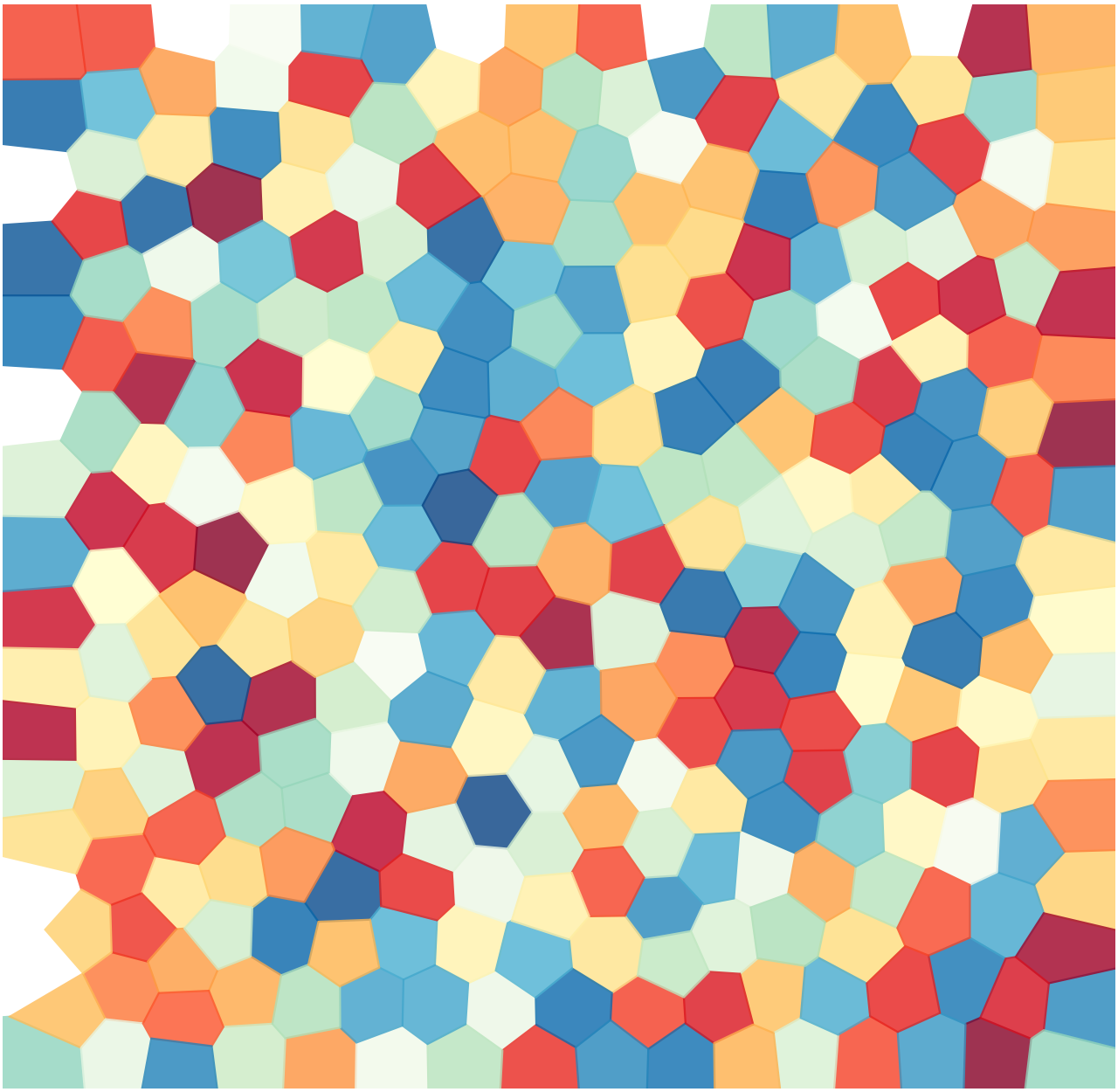
$$D\Big(\sum_i \frac{p_T}{\alpha_i}\delta_x + \mathcal{E}_1, \mathcal{E}_2\Big) = D\Big(p_T\delta_x + \mathcal{E}_1, \mathcal{E}_2\Big)$$

if $\sum_i \alpha_i = 1$. In words, splitting a single particle into multiple with the same total energy should not change the distance.

Any distance that satisfies IRC safety is said to metrize the underlying energy flow. It is easy to see from the definitions above that both exact OT and entropic OT satisfy IRC safety. More recently, a large number of jet observables, pileup mitigation techniques, and reconstruction algorithms have been explicitly cast as minimizations with respect to the EMD in ref. [71]. It is beyond the scope of this work to enumerate all the applications of optimal transport in the jet physics literature; however, we remark that the surprising breadth of both theoretical and experimental overlap is a motivation for extending this line of inquiry.

Part II

Theoretical improvements



Transport-style distances have been historically underused in machine learning due to their computational complexity, in both memory and runtime. In addition, standard optimal transport lacks some properties that are desirable in certain applications – it is not end-to-end differentiable, making it difficult to use as a loss function in a neural network; it does not induce a kernel, meaning it cannot be used in SVMs or kernel density estimators; and, the metric space associated with it is not flat in high dimensions, so standard notions of Euclidean geometry do not carry over. However, recent developments have made transport more appealing. For example, a transport-based distance known as the sliced Wasserstein distance has been shown to induce a positive definite kernel [72]. As many anomaly detection techniques are kernel-based, this suggests a potential avenue for using transport to find anomalies in collider data. Our goal in this part is to further develop the theory behind transport-based kernels and how to estimate, approximate, and compute them in practice.

This part will discuss some interesting theoretical results concerning Wasserstein distances, kernels, and approximations thereof. Chapter 4 will introduce exactly what a kernel is and why it has played such a central role in the history of machine learning, before diving into a new result on approximating kernels. and more generally, bounding the Monte Carlo quadrature error, using centroids as the empirical approximation. Chapter 5 is based on an unpublished work and is mainly devoted to speed. This chapter focuses primarily on how to make computing transport distances fast while maintaining theoretical guarantees.

Chapter 4

Kernels and approximations

> "Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise."
>
> —John Tukey

The kernel trick is an old and important technique in machine learning, used to provide additional expressivity to models for a wide range of downstream tasks [73]. A kernel is, simply, a nonlinear function that measures the similarity between two objects. Kernels are useful because they induce an embedding of some datapoint into a higher-dimensional space. Therefore, a kernelized model can leverage the expressivity of this space just through computations of the kernel, without constructing the embedding explicitly. Kernels are often used in anomaly detection models [74, 75], and are especially useful when the datapoints $\mathcal{D}$ are not drawn from Euclidean space. For example, in our case, particle physics events are empirical measures, and, as we will see in the next section, we can create a kernel so that the feature space is Euclidean. This will allow us to modify existing anomaly detection techniques by first applying our kernel to the particle physics datasets. Therefore, the techniques we outline in this section are useful to establish the underlying geometry for our anomaly detection techniques.

In this section, we will first introduce the concept of a kernel more rigorously, and state some important results about the relationship between the algebraic properties

of the kernel function and the geometric properties of the feature space it induces. Next, we will also discuss how to quickly compute or approximate the kernel matrix for a dataset. A major problem with kernel methods is that they scale quadratically-or-worse in the number of datapoints, making them impractical for extremely large datasets. To this end, we introduce the Nyström method, which is a technique for approximating the full kernel matrix by column sampling. Finally, our main contribution is to show that sampling the columns so that they are approximately the centroids of the dataset will provide an optimal approximation. Specifically, we will demonstrate a new bound on the quadrature error of an arbitrary function with bounded Hessian, when integrating across the empirical measure induced by the centroids. We leverage a relationship between Voronoi tessellations and centroids, and along the way we connect our analysis to several common machine learning techniques like k-means and Wasserstein barycenters. Later in this work, the approximation techniques we outline in this section will be used to make anomaly detection methods faster and more scalable to large datasets.

4.1 Kernels, Hilbert spaces, and all that

4.1.1 Preliminaries

Definition 4.1.1. *(Kernel function.)* A kernel is a nonlinear measure of similarity. Specifically, given a pair of datapoints $(\mathbf{x}, \mathbf{y}) \subset \mathcal{X} \times \mathcal{X}$, a kernel is defined as:

$$k(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle_{\mathcal{V}}$$

where $\mathcal{V}$ is a high-dimensional vector space and $\Phi : \mathcal{X} \to \mathcal{V}$ is a nonlinear embedding. A kernel is *positive definite* if the following condition holds:

$$\sum_{i,j}^{n} c_i c_j k(\mathbf{x}_i, \mathbf{x}_j) \geqslant 0$$

for all $\{\mathbf{x}_i\}_{i=1}^{n} \subset \mathcal{X}$ and any vector $\mathbf{c} \in \mathbb{R}^n$.

If a kernel is positive definite, it induces a space of functions $f : \mathcal{X} \to \mathbb{R}$ called the *reproducing kernel Hilbert space.*

Definition 4.1.2. *(Reproducing kernel Hilbert space.)* The Hilbert space $\mathcal{H}$ is the set of functions $f : \mathcal{X} \to \mathbb{R}$ induced by a kernel. It satisfies:

$$k(\mathbf{x}, \cdot) \in \mathcal{H}$$
$$f(\mathbf{x}) = \langle f, k(\mathbf{x}, \cdot) \rangle.$$

If a kernel is reproducing, then it is not necessary to ever construct the feature map explicity to recover the value of the kernel. In particular, given a distance or similarity measure between two datapoints, there exist certain transformations that induce a kernel in a feature space that is infinite dimensional. One such common family of kernels is the *Gaussian kernel*, which is defined as follows:

$$k_\gamma(\mathbf{x}, \mathbf{y}) = e^{-\gamma \cdot d(\mathbf{x}, \mathbf{y})}$$

for some distance function $d(\cdot, \cdot)$. Even if $\mathcal{X} = \mathbb{R}^d$ for a finite $d < +\infty$, the induced feature map $\Phi(\mathbf{x})$ maps onto an infinite dimensional feature space $\mathcal{V}$. To prove this, it suffices to Taylor expand the exponential. Next, we will state two results that apply for the Gaussian kernel. First, we show that the Gaussian kernel can be expanded in terms of an infinite series of eigenfunctions:

Theorem 4.1.1 (Mercer's theorem)**.** *Let* $L_2(\mathcal{X}, \mu)$ *be the space of* μ*-square-integrable functions on* $\mathcal{X}$ *with respect to some measure* μ*. Define the integral operator* $T_k : L_2(\mathcal{X}, \mu) \to L_2(\mathcal{X}, \mu)$ *as:*

$$(T_k f)(\cdot) \int_{\mathcal{X}} k(\cdot, x) f(x) d\mu$$

Any positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *can be eigen-decomposed as:*

$$k(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} \lambda_i \psi_i(\mathbf{x}) \psi_i(\mathbf{y})$$

where $\psi_i(\cdot)$ *are the eigenfunctions of the integral operator.*

As a consequence of this theorem, the feature map $\Phi(\mathbf{x})$ has the explicit decomposition

$$\Phi(\mathbf{x}) = (..., \sqrt{\lambda_i}\psi_i(\mathbf{x}), ...)^T$$

and, for the case of the Gaussian kernel, $\Phi \in \mathcal{C}^\infty(\mathcal{X})$, meaning it is infinitely differentiable. Finally, we define the notion of *separability* in Hilbert spaces.

Definition 4.1.3. A Hilbert space is separable if it has a countable basis.

The Hilbert space induced by the Gaussian kernel is separable [76].

4.1.2 Topology of the feature space

It is natural to next ask what a kernel can tell us about the topology and metric of the base space over which it is defined. In most cases, the answer is (unfortunately) very little. However, if the kernel is Gaussian, and it is also positive definite, it can be shown that the underlying metric space must be flat [77]. More rigorously, the theorem can be stated as follows:

Theorem 4.1.2 (Feragen, 2015)**.** *Given a Riemannian manifold with metric tensor* $(\mathcal{M}, g)$*, with a line element given as* $ds^2 = g_{\mu\nu}\mathbf{x}^\mu\mathbf{x}^\nu$*, the manifold is Euclidean if and only if the kernel*

$$k(\mathbf{x}, \mathbf{y}) = e^{-\gamma d^2(\mathbf{x},\mathbf{y})}$$

is positive definite for all values of γ*.*

For the remainder of this work, we will concern ourselves primarily with kernels that are both Gaussian and positive definite. In this case, the following additional fact, due to ref. [76] will be useful.

Remark. *Assume that the space* $(\mathcal{X}, d)$ *is isometrically Euclidean, and* k *is the Gaussian kernel induced by* d*. Then, for any subset* $\mathcal{A} \subset \mathcal{X}$*, the image of the subset in the RKHS of the kernel* k *its itself a flat manifold, with metric tensor given by* $g'_{\mu\nu} = \gamma \cdot \delta_{\mu\nu}$*.*

This result is a corollary of Nash's theorem, which states that any Riemannian manifold can be isometrically embedded in (higher-dimensional) Euclidean space. It holds for any positive definite kernel that induces a separable reproducing kernel Hilbert space. Intuitively, this remark states that the similarity defined by the feature map induced by Gaussian kernel satisfies many of our standard notions of Euclidean geometry. We will return to this result in the next section.

4.2 Nyström's approximation

In many kernel-based algorithms, the key object of study is the Gram matrix $\mathbf{K}$ of a dataset $\mathcal{D}$, which is the symmetric, positive semi-definite matrix defined as $\mathbf{K}_{ij} = \mathsf{k}(\mathbf{x}_i, \mathbf{x}_j)$. Unfortunately, this causes issues in the case where $\mathsf{k}(\cdot,\cdot)$ is expensive or when the cardinality of the dataset $n = |\mathcal{D}|$ is large. In some cases, there are shortcuts to approximate this matrix. When the kernel is translation invariant, it can be decomposed into a convex combination within the cone of Fourier kernels. By performing a Monte Carlo approximation of the Fourier transform of the kernel, one can find a low-rank approximation of the full kernel matrix, known as the Random Fourier Feature method [78]. A popular alternative is the Nyström method, which approximates the Gram matrix through an eigenvalue decomposition [79, 80].

Definition 4.2.1. *(Nyström method.)* Given a positive semidefinite matrix

$$\mathbf{K} = \begin{pmatrix} \mathbf{A} & \mathbf{B}^\intercal \\ \mathbf{B} & \mathbf{C} \end{pmatrix}$$

where $\mathbf{A} \in \mathbb{R}^{m \times m}$, $\mathbf{B} \in \mathbb{R}^{m \times n-m}$, the Nyström approximation is

$$\tilde{\mathbf{K}} = \begin{pmatrix} \mathbf{A} & \mathbf{B}^\intercal \\ \mathbf{B} & \mathbf{B}^\intercal \mathbf{A}^{-1} \mathbf{B} \end{pmatrix} \tag{4.1}$$

Applying the Nyström technique on a given Gram matrix then reduces to sampling

a subset $\mathcal{A} \subseteq \mathcal{D}$ of datapoints (known as "landmarks"), computing the distances from each landmark to the rest of the dataset, and applying the formula in Eq. (4.1). In particular, it can be shown that this approximation is simply a truncation of the explicit feature map given by Mercer's theorem.

To show this, note that the integral operator $\mathcal{T}_k$ can be approximated by Monte Carlo quadrature against the measure μ as follows:

$$\begin{aligned}(\mathcal{T}_k \phi_i)(\mathbf{x}) = a_i \phi_i(\mathbf{x}) &= \int_{\mathcal{X}} k(\mathbf{x}, \mathbf{y}) \phi_i(\mathbf{x}) d\mu \\ &= \mathbb{E}_{\mathbf{x}' \sim \mu} k(\mathbf{x}, \mathbf{x}') \phi_i(\mathbf{x}') \\ &\approxeq \frac{1}{t} \sum_{j=1}^{t} k(\mathbf{x}, \mathbf{x}_j) \phi_i(\mathbf{x}_j)\end{aligned}$$

where $\{\mathbf{x}_j\}_{i=1}^{n} = \mathcal{D}$ is the dataset. This directly leads to the symmetric eigendecomposition $\mathbf{K} = \mathbf{\Psi \Lambda \Psi}^\top$, where $\mathbf{\Psi}, \mathbf{\Lambda}$ are the eigenvectors and diagonal eigenvalue matrix, respectively. In the Nyström technique, the singular value decomposition above is approximated using a subset of landmark points $\{\mathbf{z}_j\}_{j=1}^{m} = \mathcal{Z}$. The optimal approximation is given by:

$$\begin{aligned}\mathbf{\Psi}_{\mathcal{D}} &\approxeq \sqrt{\frac{m}{n}} \mathbf{K}_{\mathcal{D},\mathcal{Z}} \mathbf{\Psi}_{\mathcal{Z}} \mathbf{\Lambda}_{\mathcal{Z}}^{-1} \\ \mathbf{\Lambda}_{\mathcal{D}} &\approxeq \frac{n}{m} \mathbf{\Lambda}_{\mathcal{Z}},\end{aligned}$$

where $\mathbf{K}_{\mathcal{D},\mathcal{Z}}$ is the matrix whose entries are $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{z}_j)$. Plugging this approximation in to reconstruct $\mathbf{K}$ gives exactly the formula in Eq. (4.1) This technique is appealing not only for its theoretical guarantees, but also for its computational advantage – if the number of landmarks chosen is k, then the computational complexity is $\mathcal{O}(k^2 n)$, significantly smaller than the full-rank $\mathcal{O}(n^3)$.

Choosing the set of landmarks for the Nyström technique intelligently can yield improved theoretical and practical approximations. Techniques such as determinantal point processes, k-means, and k-means++ have been utilized as initialization steps to

select good landmark points at the expense of some additional precomputation [81–83]. In particular, the analysis presented in ref [81], shows that k-means clustering in feature space is equivalent to selecting the top k eigenvectors in the SVD step, meaning that this technique leads to a good approximation of the Gram matrix in terms of Frobenius norm. In our work, we will focus instead on bounding the error of quadrature of some integral under an empirical approximation induced by the kernel landmarks.

4.2.1 Quadrature and Voronoi tessellations

Recent work has attempted to bound the quadrature error of Nyström-type approximations for arbitrary kernels [84–86]. However, most results either appeal to a specific learning paradigm or apply quadrature in the RKHS. By contrast, our analysis will focus on quadrature bounds in the feature space directly. In this section, we will focus primarily on k-means initializations, and, in particular, their relation to centroidal Voronoi tessellations and low-discrepancy sequences [87]. First, we will define some important terms.

Definition 4.2.2. *(Centroidal Voronoi tessellation).* Assume we are given a measure μ on a metric space $(\mathcal{X}, d)$. The Voronoi tessellation induced by a set of points $\mathcal{Z} = \{\mathbf{z}_i\}_{i=1}^k$ is defined as:

$$\mathcal{V} = \left\{ V_i \,|\, V_i = \{\mathbf{x} \in \mathcal{X} : \min_{\mathbf{z} \in \mathcal{Z}} \|\mathbf{x} - \mathbf{z}\| = \mathbf{z}_i\} \right\}$$

A tessellation is known as a centroidal Voronoi tessellation if, in addition to the property above, the induced points are the centroids of the respective cells – that is:

$$\mathbf{z}_i = \frac{\int_{V_i} \mathbf{x} d\mu}{\int_{V_i} d\mu}$$

The centroidal Voronoi tessellation is the minimizer of a specific energy function.

We define the *CVT energy* for a set of landmarks as

$$\mathsf{E}(\mathcal{Z}) = \sum_{i} \int_{V_i} \|\mathbf{x} - \mathbf{z}_i\|^2 \mathrm{d}\mu \tag{4.2}$$

Algorithms to minimize the CVT energy are also good approximations to optimal solution to k-means problem. In fact, the CVT energy and the k-means "within-cluster-sum-of-squares" error are equivalent. In particular, we point to the algorithms due to Lloyd and MacQueen [88], which are guaranteed to converge to a centroidal Voronoi tessellation, and also are commonly used to solve the k-means problem. Finally, the k-means++ initialization scheme provides a fast approximation to the optimal objective with only $\mathcal{O}(\log k)$ penalty [81].

We note that, when the measure μ is not the Lebesgue measure, this energy has a close relationship to semidiscrete optimal transport [89]. In particular, the optimal transport value can be written as [90]:

$$W_2^2\left(\mu, \frac{1}{k}\sum_{i=1}^{k} \delta(\mathbf{z}_k)\right) = \sup_{\varphi} \sum_{i=1}^{k} \left(\frac{1}{k}\varphi^i + \int_{V_\varphi^i} (\|\mathbf{x} - \mathbf{z}_i\|^2 - \varphi^i)\mathrm{d}\mu\right)$$

where $\varphi \in \mathbb{R}^k$ is the Kantorovich potential. If $\int_{V_\varphi^i} \mathrm{d}\mu = \frac{1}{k}$, then the semidiscrete 2-Wasserstein distance is equal to the CVT energy. We will see later that, for an optimal CVT, this is indeed the case. Further, it has recently been shown that finding the Wasserstein barycenter (i.e., the distribution supported on k points that most closely approximates the target μ in Wasserstein distance) is equivalent to a centroidal power tessellation [90].

Our goal is to show that selecting a subset of points according to the centroidal Voronoi tessellation scheme (or approximations thereof) leads to provable bounds on the quadrature of arbitrary functions. To proceed, we take inspiration from ref. [91], and generalize their argument to arbitrary measurable spaces. In particular, assume that the data $\mathcal{D}$ is drawn from some measure μ, and our target function $f \in \mathcal{C}^\infty(\mathcal{X})$

is Lipschitz with constant L. In this case, the quadrature error can be written:

$$\mathsf{Err}(\mathcal{Z}) = \left\| \int_{\mathcal{X}} f(\mathbf{x}) d\mu - \frac{1}{k} \sum_{i=1}^{k} f(\mathbf{z}_i) \right\| \tag{4.3}$$

Taylor expansion of the first integral yields the decomposition:

$$\int_{\mathcal{X}} f(\mathbf{x}) d\mu \lesssim \sum_i \left[\int_{V_i} f(\mathbf{z}_i) + \nabla f(\mathbf{z}_i)^\intercal (\mathbf{x} - \mathbf{z}_i) + \frac{1}{2} (\mathbf{x} - \mathbf{z}_i)^\intercal (\nabla^2 f(\mathbf{z}_i))(\mathbf{x} - \mathbf{z}_i) d\mu \right] \tag{4.4}$$

$$= \sum_i \left[\mu(V_i) f(\mathbf{z}_i) + \int_{V_i} \nabla f(\mathbf{z}_i)^\intercal (\mathbf{x} - \mathbf{z}_i) + \int_{V_i} \frac{1}{2} (\mathbf{x} - \mathbf{z}_i)^\intercal (\nabla^2 f(\mathbf{z}_i))(\mathbf{x} - \mathbf{z}_i) d\mu \right] \tag{4.5}$$

$$\leqslant \sum_i \mu(V_i) f(\mathbf{z}_i) + \sup_{\mathbf{a}:\|\mathbf{a}\|=1} \|\mathbf{a}^\intercal \nabla^2 f(\mathbf{z}_i) \mathbf{a}\| \int_{V_i} \|\mathbf{x} - \mathbf{z}_i\|^2 d\mu \tag{4.6}$$

where ∇ is the gradient, $\mathbf{H} = \nabla^2$ is the Hessian, and the middle term vanishes (and is minimal) if $\mathbf{z}_i$ is the centroid of V_i. Assuming that the target function f is convex and twice differentiable, the operator norm of the Hessian is bounded above by L. This is simply due to the mean value theorem and the fact that the Hessian is positive semidefinite for any convex and differentiable function f:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x} + \Delta\mathbf{x})\| \leqslant L \|\Delta\mathbf{x}\| \iff \|\nabla^2 f\| \leqslant L$$
$$\implies \mathbf{a}^\intercal (\nabla^2 f(\mathbf{x})) \mathbf{a} \leqslant L \mathbf{a}^\intercal \mathbf{a} \quad \forall \mathbf{a}$$

Plugging this into eq. (4.3) gives the bound:

$$\mathsf{Err}(\mathcal{Z}) \leqslant \sum_i f(\mathbf{z}_i) \left| \mu(V_i) - \frac{1}{k} \right| + \sup_{\mathbf{a}} \|\nabla^2 f(\mathbf{a})\| \cdot \sum_i \int_{V_i} \|\mathbf{x} - \mathbf{z}_i\|^2 d\mu \tag{4.7}$$

$$= \sum_i f(\mathbf{z}_i) \left(\mu(V_i) - \frac{1}{k} \right) + L \cdot \mathsf{E}(\mathcal{Z}) \tag{4.8}$$

Finally, we note a conjectured bound on the optimal CVT error, which has been

proven for two and three dimensions [92].

Theorem 4.2.1. (Gersho's conjecture.) *As* $k \to \infty$, *the optimal CVT (i.e., the one that minimizes the CVT energy) is a tessellation such that the following two conditions hold:*

1. *All Voronoi cells* V_i *are geometrically similar to some reference polytope* V, *which depends on the dimension.*

2. *Each* V_i *has equal measure with respect to* μ. *Specifically,*

$$\mu(V_i) \equiv k^{-1/d} \cdot \mathsf{vol}(V)$$

where vol *is the volume with respect to the Lebesgue measure,* d *is the dimension and* k *is the number of points in the CVT. Further, these conditions hold for arbitrary* μ.

For 2 dimensions, Gersho's conjecture can be visualized in Fig. 4-1. Note that the optimal polytope for $\mathbb{R}^2$ is a hexagon, and that the regularity of the tessellation increases from the top to the bottom of the figure. Following [91], we see that the following quantity is invariant to the deformations permitted under Gersho's conjecture:

$$M_i = \frac{\int_{V_i} \|\mathbf{x} - \mathbf{z}_i\|^2 d\mu}{\left(\int_{V_i} d\mu\right)^{(d+2)/d}}$$

If Gersho's conjecture is valid, then all $M_i = M$ for some optimal tessellation $\mathcal{V}$. Therefore:

$$\begin{aligned} E(\mathcal{Z}) &= \sum_i \int_{V_i} \|\mathbf{x} - \mathbf{z}_i\|^2 d\mu \\ &= \sum_i \left(\int_{V_i} d\mu\right)^{(d+2)/d} M_i \\ &\approxeq k^{-2/d} M = \mathcal{O}(k^{-2/d}) \end{aligned}$$

In fact, this bound is reminiscent of the best asymptotic rate for an empirical

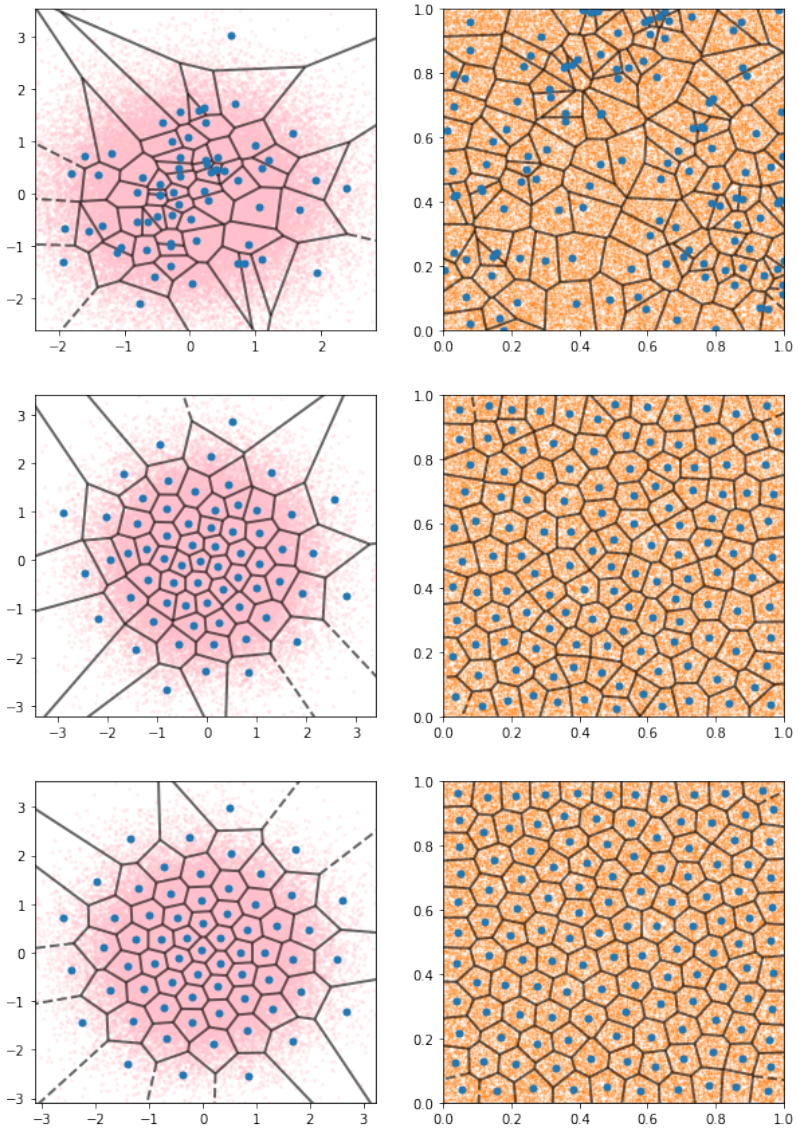


Figure 4-1: A visualization of Voronoi tessellations in 2 dimensions for a Gaussian (left) and uniform (right) distribution. From top to bottom, the landmarks are chosen randomly, using k-means++, and using Lloyd's algorithm.

approximation in Wasserstein distance, which is $\mathcal{O}(k^{-1/d})$, independent of the points sampled. The similarity may follow, intuitively, from the idea that the Wasserstein distance metrizes weak convergence of the empirical measure. Gersho's conjecture also suggests that the best subset to select to minimize quadrature error is the CVT. This follows from the two terms in Eq. (4.7). By Gersho's conjecture, the first term vanishes, and the second term is minimized by the definition of the CVT.

Putting all the arguments above together, this implies that:

$$\boxed{\left\| \int_{V_i} f(\mathbf{x}) d\mu - \frac{1}{k} \sum_i f(\mathbf{z}_i) \right\| = \mathcal{O}(k^{-2/d})}$$

if $\mathbf{z}_i$ are the centroids of an optimal Voronoi tessellation. A similar bound holds if the points are, instead, approximations to a CVT – for example, in k-means++, the quadrature error will have an additional $\mathcal{O}(\log k)$ term. This provides a powerful bound on an optimal method of sampling to minimize the quadrature error of arbitrary functions.

4.2.2 Centroids as landmarks

To conclude this section, we will connect back the result we have shown to the Nyström approximation. To make the relationship between these two techniques explicit, consider a generalized kernel method, like the support vector machine. The output function in a kernelized SVM is:

$$\widehat{y}(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} w_i y_i k(\mathbf{x}, \mathbf{x}_i) \right)$$

for some weight vector $\mathbf{w}$ and kernel function $\kappa(\cdot, \cdot)$. Naively, computing the prediction requires the full kernel matrix $\mathbf{K}$. However, the Nyström approximation suggests that instead of computing the prediction with respect to all points in the dataset, we can sample a subset $\{\mathbf{z}_j\}_{j=1}^{k}$ by simply selecting points randomly. However, the arguments above, applied with care, suggest that this error can be minimized when the subset is a CVT. First, we note a problem with directly applying our results

under the feature map. Note that, due to Theorem 4.1.2, all the arguments above hold equivalently under the transformation by the feature map of a positive definite kernel. In particular, this means that computing an approximate Voronoi tessellation in the data space gives a similar tessellation in the feature space. However, this comes with the caveat that the centroids are almost certainly not an element of the original dataset – i.e., $\mathbf{z}_i \notin \mathcal{D}$. This means that, without explicitly constructing the feature map, it will (in general) not be possible to achieve the tight bound described above. Further, the dimension of the feature space is not guaranteed to be finite, and the subspace induced by the image of the dataset $\mathsf{im}(\mathcal{D})$ under the map Φ can have very high dimension, so the error bounds may be very loose.

Instead, let us consider the approximation of the kernel function itself. Assume that the kernel κ is Gaussian with parameter γ. In particular, note that the quadrature error bound when applied to the function $f(\mathbf{x}) = \sum_{i=1}^{N} \kappa(\mathbf{x}_i, \mathbf{x})$ over the landmarks $\{\mathbf{z}_j\}_{j=1}^{k}$ gives the following:

$$\begin{aligned}
\left\| \sum_{i=1}^{N} \int_{\mathcal{X}} \kappa(\mathbf{x}_i, \mathbf{x}) \, d\mu - \frac{1}{k} \sum_{i=1}^{N} \sum_{j=1}^{k} \kappa(\mathbf{x}_i, \mathbf{z}_j) \right\| &\leqslant \sum_{i=1}^{N} \left\| \int_{\mathcal{X}} \kappa(\mathbf{x}_i, \mathbf{x}) \, d\mu - \frac{1}{k} \sum_{i=1}^{N} \sum_{j=1}^{k} \kappa(\mathbf{x}_i, \mathbf{z}_j) \right\| \\
&\leqslant \frac{1}{k} \sum_{i=1}^{N} \sum_{j=1}^{k} \left\| \nabla^2 \kappa(\mathbf{z}_j) \right\| \cdot \int_{V_i} \|\mathbf{x} - \mathbf{z}_j\|^2 d\mu \\
&\leqslant \frac{1}{k} \sum_{i=1}^{N} \sum_{j=1}^{k} -\gamma \cdot \int_{V_i} \|\mathbf{x} - \mathbf{z}_j\|^2 d\mu
\end{aligned}$$

where we have used the fact that the Hessian of the Gaussian kernel is the inverse of its covariance matrix. This value is, by the same argument as above, minimized when the landmarks are an optimal CVT, and additionally, the bound holds with d the dimension of the base space $\mathcal{X}$.[1]

[1]This also suggests, interestingly, that a distribution can be well approximated with respect to the *kernel mean map* with a subset of points equivalent to the CVT. Similarly, the CVT is the best empirical distribution to approximate the true measure with respect to the Wasserstein distance. We will define the kernel mean map in the next section, but implementing this approximation is

While we leave a full analysis to future work, a similar argument can also be applied to functionals of the kernel – i.e., consider the mean squared error $\mathsf{MSE} = \sum_i (\mathsf{y}(\mathbf{x}_i) - \widehat{\mathsf{y}}(\mathbf{x}_i))^2$. A future analysis can hopefully show that the objective attained when minimizing this functional over the weight vector $\mathbf{w}$ supported at only the points in the subset is close to that achieved when solving the quadratic program fully over the whole dataset.

4.3 Conclusion

Kernels are a useful tool in machine learning to improve the expressivity of a model. In particular, kernels allow us to use non-linear embeddings into potentially infinite-dimensional feature spaces without explicitly constructing the embedding map. Unfortunately, most kernel-based methods suffer from quadratic-or-worse time and space complexity due to the burden of computing the full Gram matrix. As an example, the anomaly detection techniques we will discuss in Chapter 6 both are kernel-based. Therefore, it is often necessary to employ approximations to make these methods practical. In this section, we have introduced the Nyström method, a low-rank approximation technique based on column sampling. We provide what is, to our knowledge, a novel analysis of this technique for positive definite kernels that have bounded Hessian by framing the error in terms of quadrature of an arbitrary function on the Gram matrix. Using this method, we have shown that there are theoretical benefits, in terms of this error, to initializing the landmarks of this technique with the centroids of the measure. We connect this technique to recent developments in optimal quantization, specifically the semidiscrete Wasserstein barycenter problem.

Before we apply centroid-based kernel approximations to our anomaly detection techniques, however, we still need a kernel that operates on distributions. We would like this kernel to satisfy the properties of positive definiteness while still respecting the notion of weak convergence as described in Chapter 5. As we will see in the next chapter, the standard Gaussian kernel defined over the Wasserstein distance does

outside the scope of this work.

not provide these theoretical guarantees, except in 1 dimension. Therefore, we will introduce and show how to compute a variant of the Wasserstein distance based on projecting the atoms of the distribution to $\mathtt{d} = 1$.

Chapter 5

Fast Wasserstein-type kernels

"Numerical analysis is quickly becoming an experimental science."

— Peter Wynn

In this section, we will describe our recent progress in constructing fast, unbiased estimators for transport-related distances, and their applicability in kernel-based machine learning techniques. Recall that the central object of study in this thesis is the event signature, which is an empirical distribution. To operate on these distributions in a practical setting, it is convenient to first map them into a Hilbert space defined by some kernel. In that space, one can measure lengths, angles, and local densities, which are important in many machine learning techniques, including (but not limited to) models of anomaly detection. To this end, certain kernels operating on distributions [72, 93] have been proposed. We will specifically analyze one kernel, the *sliced Wasserstein* kernel, and show that the naïve estimator proposed in the literature for it suffers from bias with respect to the number of points in the empirical distribution. Because of this bias, the theoretical benefits it enjoys are not necessarily guaranteed in practice.

In the first part of this chapter, we will define the sliced Wasserstein distance, and provide an unbiased estimator for it. We show both numerically and mathematically that our estimator eliminates the bias inherent to the naïve estimator. We will also show that, in the case that the empirical distribution is an approximation to some absolutely continuous measure, that the "bootstrapped" version of the kernel generated

from the empirical distributions also satisfies kernel properties with respect to the underlying continuous measures. Following that, we will introduce another estimator for the sliced Wasserstein distance, based on a multi-level Monte Carlo scheme. This estimator is optimized for speed, and we will prove a theoretical runtime complexity for it that is lower than that of the unbiased and naïve estimators. This chapter is primarily based on work done jointly with Justin Solomon, Sam Power, Abdelkader Baggag, and Yue Wang.

5.1 Sliced Wasserstein distances

In one dimension, the Wasserstein distance has a closed form and is easy to compute. The family of *sliced Wasserstein distances* exploits this computational advantage, by computing the Wasserstein distance between 1-dimensional projections onto random vectors. More precisely, it can be defined as follows:

Definition 5.1.1. Let $\mathbb{S}^{n-1}$ denote the unit sphere in $\mathbb{R}^n$. Then, the *sliced Wasserstein* distance between $\mu, \nu \in \mathcal{P}_+(\mathbb{R}^n)$ equals:

$$\mathsf{SW}^2(\mu, \nu) := \mathbb{E}_{\theta \sim \mathbb{S}^{n-1}}[\mathcal{W}_2^2(\mathsf{proj}_\theta\, \mu, \mathsf{proj}_\theta\, \nu)]. \tag{5.1}$$

In this expression, $\mathsf{proj}_\theta\, \mu$ denotes the projection of each atom in measure μ onto the vector θ. In words, Eq. (5.1) is the expected transport distance between the projections of μ and ν onto arbitrary lines through the origin.

The sliced Wasserstein distance has been utilized in many fields as a replacement for the Wasserstein distance. For example, as it can be easily differentiated, it can be used as a loss function for training neural networks with backpropagation [94, 95]. In this form, it has appeared in natural language processing [95], music transcription [96], and computer vision [97, 98], among others. In addition, its kernel-related properties have been exploited for kernel-based machine learning [72] and topological data analysis [99]. However, as we will note later, these kernel methods suffer from a bias, which we will describe and correct in this work.

While the sliced Wasserstein distance is a completely different metric to the Wasserstein, it enjoys some theoretical and computational advantages that make it appealing for use in many kernel-based problems.

- **Computational advantage.** As the integral over the sphere can be approximated stochastically with a Monte Carlo sum, an estimator for the sliced Wasserstein distance can be given as follows:

$$\widehat{\mathsf{SW}}^2(\mu, \nu) = \frac{1}{T} \sum_{k=1}^{T} \mathcal{W}_2^2(\mathsf{proj}_{\theta_k} \mu, \mathsf{proj}_{\theta_k} \nu), \tag{5.2}$$

  where $\theta_1, \ldots, \theta_T \sim \mathbb{S}^{n-1}$ are drawn i.i.d. As described in Chapter 3, 1-dimensional transport distances between empirical distributions can be computed in closed-form by sorting [45]. This means that the runtime of the above estimator is $\mathcal{O}(Tn \log n)$, much faster than the quadratic-or-worse scaling of Sinkhorn and the exact OT solver.

- **Positive definite kernel.** Unlike the regular Wasserstein distance, the sliced Wasserstein distance generates a positive definite Gaussian kernel. We can construct it from SW by defining:

$$\mathcal{K}_\gamma(\mu, \nu) = e^{-\gamma \mathsf{SW}^2(\mu, \nu)}. \tag{5.3}$$

  $\mathcal{K}_\gamma$ is positive definite for all $\gamma > 0$ on the space of absolutely continuous probability measures [72]. This property enables $\mathcal{K}_\gamma$ to be used in any standard kernel-based machine learning algorithm, whereas the non-sliced functional $(\mu, \nu) \mapsto \exp(-\gamma \mathcal{W}_2^2(\mu, \nu))$ does *not* satisfy the positive definite condition needed to do so when $n > 1$. By a similar argument as outlined in the previous chapter, this additionally implies that the space of measures endowed with the sliced Wasserstein distance is isometric to Euclidean space, and the Hilbert space induced by this kernel is separable.

Further, we note that the sliced Wasserstein distance inherits the properties of

IRC safety and energy flow metrization discussed in Chapter 3.

5.1.1 Debiasing the kernel estimator

A naive computation of the kernel as the exponential of the simple estimator in Eq. (5.2) immediately encounters an issue, however. By Jensen's inequality,

$$\mathbb{E}[\exp(-\gamma x)] \geqslant \exp(\mathbb{E}[-\gamma x])$$

which means that simply exponentiating our Monte Carlo approximation is an overestimate of the true value of the kernel.

Next, we will explicitly define our unbiased estimator. To start, we leverage the following identity:

$$1 = \mathbb{E}_K\left[\frac{\mathbb{1}(K \geqslant k)}{\mathbb{P}(K \geqslant k | K \sim \rho)}\right]$$

where $\rho \in \mathcal{P}_+(\mathbb{N})$ is a distribution defined over the natural numbers. Intuitively, this identity allows us to . Next, we will consider the Taylor expansion of the exponential, following the work of refs. [100, 101].

$$\exp(-\gamma x) = \sum_{k=0}^{\infty} \frac{(-\gamma x)^k}{k!}$$

Multiplying the equations above, and moving the sum through the expectation, we arrive at the following expression for an unbiased estimator:

$$\begin{aligned} e^{-\gamma x} &= \mathbb{E}_{K\sim \mathsf{Geom}(p)}\left[\sum_{k=0}^{\infty} \frac{(-\gamma x)^k}{k!} \frac{\mathbb{1}(K \geqslant k)}{(1-p)^k}\right] \\ &= \boxed{\mathbb{E}_{K\sim \mathsf{Geom}(p)}\left[\mathbb{E}_{\{x_k\}_{k=1}^K \sim \mathcal{X}}\left[\sum_{k=0}^{K} \frac{1}{k!}\left(\frac{-\gamma}{1-p}\right)^k \prod_{\ell=1}^{k} x_\ell\right]\right]}. \end{aligned}$$

where, as the test distribution over the natural numbers, we have chosen $\mathsf{Geom}(p)$ for some parameter $p > 0$ due to its convenient closed form, and the terms x_ℓ are the

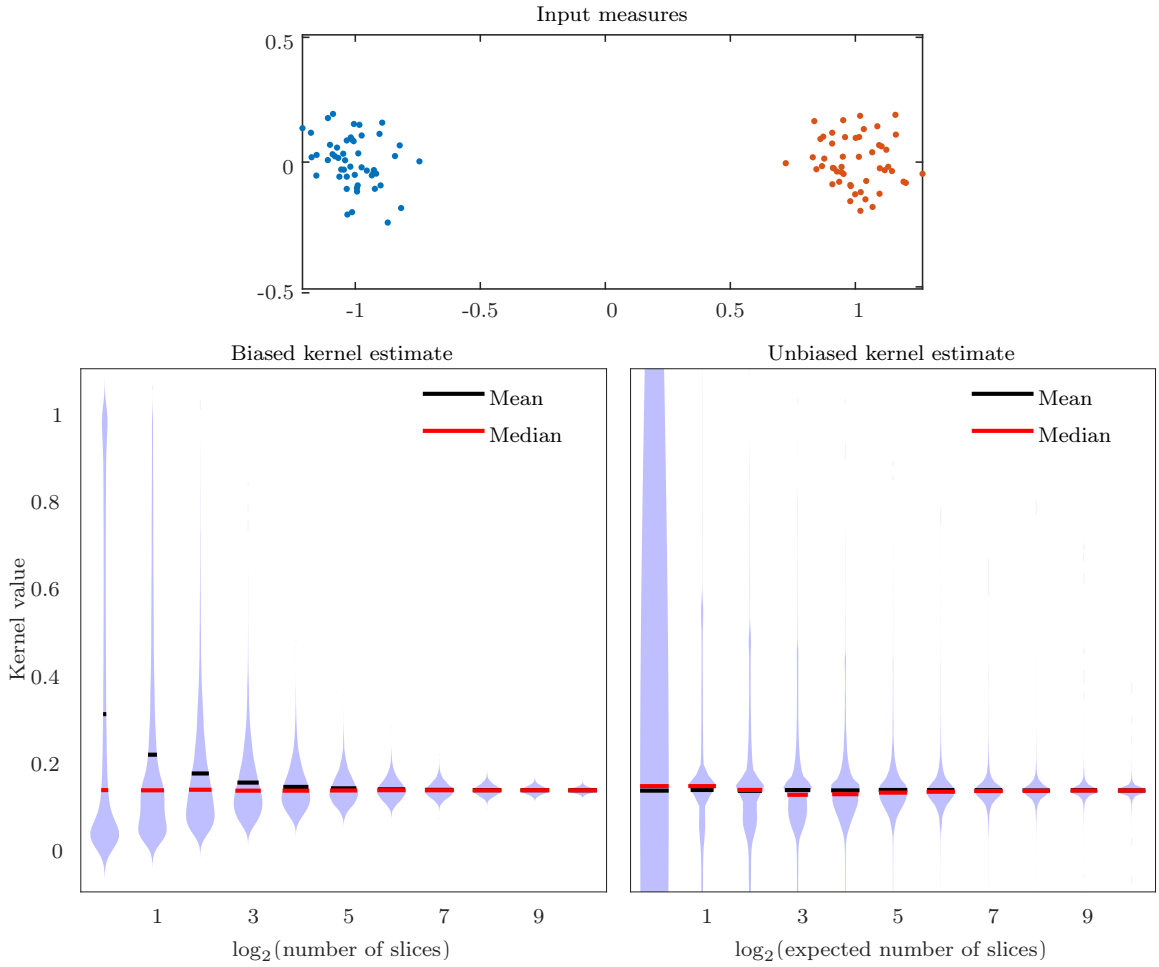


Figure 5-1: Stochastic estimation of the kernel K_γ between two 50-point empirical measures (top), with $\gamma = 1$, accumulated over 10^4 trials. The mean of the naïve biased estimate depends on the number of slices drawn (lower left), while our estimator has the same mean regardless of the number of slices (lower right).

one-dimensional transport distances computed after projecting along the slice θ_ℓ

Fig. 5-1 illustrates the bias of the naive estimator and our improvement over it in practice. When the biased kernel estimator is computed between two Gaussian measures supported on 50 points each, if the number of projections is small, the estimator suffers from a small but consistent overestimation, due to the convexity of the exponential function. In contrast, the unbiased estimator has no dependence on the number of projections, but has the drawback of any individual trial not guaranteed to fall in the range $[0, 1]$.

In practice, the unbiased estimator additionally suffers from a large variance. The variance is generated by two orthogonal sources within the computation of the unbiased estimator.

- **Taylor series truncation.** The number of terms in the Taylor series is controlled by the parameter p governing the geometric distribution. In particular,

$$\mathbb{E}_{\mathsf{K}\sim\mathsf{Geom}(\mathsf{p})}[\mathsf{K}] = \frac{1}{\mathsf{p}} - 1.$$

  For large p, this means that the expected number of terms is small, and the variance of the exponential is high.

- **Multiplying random variables.** If the variance of the underlying distribution is large with respect to the projection slices (i.e., the distribution is not isotropic), then the variance of multiplying many random variables to form higher-order terms in the Taylor series can dominate.

We will now show how to control this variance. We propose three independent modifications to the algorithm described above to reduce the variance.

Choosing the parameter p. The goal of choosing p is to select a value that is large enough for the polynomial in the numerator to be dominated by the factorial in the denominator of the last Taylor series term. In particular, this means that we

want:

$$\frac{(\gamma x)^K}{(1-p)^K K!} \lessapprox 1.$$

Substituting the expectation for K and rearranging, we find:

$$\begin{aligned}\left(\frac{\gamma x}{1 - 1/(K+1)}\right)^K &\lessapprox K! \\ &\leqslant K^K\end{aligned}$$

where x is a "typical" value for the sliced Wasserstein distance, and we have used Stirling's inequality in the second line. Taking the K-th root of both sides and noting that the denominator is always at least equal to $\frac{1}{2}$, we obtain the final heuristic for p:

$$K \approxeq 2\gamma x \implies p \approxeq \frac{1}{1+2\gamma x}$$

Avoiding reuse. Note that in our estimator, the product of the sliced distances in the interior of the expectation is asymmetric – in particular, the first terms are used much more often than the last. While this does not induce a bias in the estimator, it does increase the variance. Two easy remedies present themselves to avoid this.

1. **Random shuffling.** If we randomly permute the values before taking the product, in expectation, this problem is avoided. However, in any given iteration, we might still be "unlucky" and use one term more than any other. This scenario is more likely if the number of projections is small, which is already a scenario with higher variance.

2. **Symmetric polynomials.** Alternatively, we can replace the product with an elementary symmetric polynomial:

$$\prod_\ell x_\ell \to \binom{K}{k}^{-1} E_k(x_1, ..., x_K)$$

which is the sum of all possible k-element combinations of the K arguments $x_1, ..., x_K$. Computationally, this step takes $\mathcal{O}(K^2)$ time with a dynamic pro-

gramming algorithm, and is implemented in log domain for numerical stability.

Centering the Taylor series. Instead of centering the Taylor series at 0, we can center it instead closer to the mean, which will reduce variance in the approximation. In particular, we can subtract the mean $\bar{x} = \sum_\ell x_\ell$ from each distance, and then add this mean back at the end by multiplying through with the factor $\exp(-\gamma\bar{x})$ to recover the correct value. This computation both adds numerical stability and decreases the variance.

5.1.2 Bootstrapping the kernel

There are certain applications for which computing the sliced Wasserstein kernel between two distributions is computationally prohibitive – for example, if the number of points in each empirical distribution is extremely large, or if the distributions are absolutely continuous and sample access is costly. In this case, it is tempting to approximate the sliced Wasserstein distance with a subsampled or bootstrapped version, defined as:

$$\mathsf{SW}^2(\mu,\nu) \approx \mathbb{E}_{x_i\sim\mu y_i\sim\nu}\left[\mathsf{SW}^2\left(\frac{1}{N}\sum_{i=1}^N \delta_{x_i}, \frac{1}{N}\sum_{i=1}^N \delta_{y_i}\right)\right]. \tag{5.4}$$

However, this approximation is biased, and does not induce a distance. Consider the bootstrapped sliced Wasserstein "distance" from an absolutely continuous measure to itself. For any finite number of points, the transport distance will be positive, as the probability of the samples being the same is vanishingly small. Despite this, we can define a bootstrapped kernel as follows:

$$\overline{\mathcal{K}}_\gamma^N(\mu,\nu) := \mathbb{E}_{x_i\sim\mu y_j\sim\nu}\left[\mathcal{K}_\gamma\left(\frac{1}{N}\sum_{i=1}^N \delta_{x_i}, \frac{1}{N}\sum_{j=1}^N \delta_{y_j}\right)\right]. \tag{5.5}$$

A straightforward argument verifies that $\overline{\mathcal{K}}_\gamma^N$ is a kernel. Since $\mathcal{K}_\gamma$ is a kernel, there exists a feature map $\Phi : \mathcal{P}_+(\mathbb{R}^n) \to \mathcal{H}$ into a Hilbert space $\mathcal{H}$ with $\mathcal{K}_\gamma(\mu,\nu) =$

$\langle \Phi(\mu), \Phi(\nu) \rangle_{\mathcal{H}}$. Then,

$$\overline{\mathcal{K}}_\gamma^N(\mu,\nu) = \left\langle \mathbb{E}_{x_i\sim\mu}\Phi\left(\frac{1}{N}\sum_{i=1}^N \delta_{x_i}\right), \mathbb{E}_{y_i\sim\nu}\Phi\left(\frac{1}{N}\sum_{i=1}^N \delta_{y_i}\right)\right\rangle.$$

Hence, $\overline{\mathcal{K}}_\gamma^N$ is an inner product via a new feature map

$$\mu \overset{\tilde{\Phi}}{\mapsto} \left[\mathbb{E}_{x_i\sim\mu}\Phi\left(\frac{1}{N}\sum_{i=1}^N \delta_{x_i}\right)\right]. \tag{5.6}$$

Moreover, since sampled transport distances converge to the true value as $N \to \infty$, we can verify

$$\mathcal{K}_\gamma(\mu,\nu) = \lim_{N\to\infty} \overline{\mathcal{K}}_\gamma^N(\mu,\nu). \tag{5.7}$$

To prove this, consider the following scenario. For a sampled set of points $x_i \sim \mu \in \mathcal{P}_+(\mathbb{R}^n)$, define $\mu_N = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$. Berthet et al. [102, Theorem 12] prove in one dimension that $\lim_{N\to\infty} \mathcal{W}_p(\mu_N,\nu_N) = \mathcal{W}_p(\mu,\nu)$ almost surely. Hence, starting from (5.1) we have

$$\begin{aligned}
&\lim_{N\to\infty} \mathsf{SW}_2(\mu_N,\nu_N) \\
&= \lim_{N\to\infty} \mathbb{E}_{v\sim\mathbb{S}^{n-1}}\left[\mathcal{W}_2^2(\mathsf{proj}_v\,\mu_N, \mathsf{proj}_v\,\nu_N)\right] \\
&\overset{\text{a.s.}}{=} \mathbb{E}_{v\sim\mathbb{S}^{n-1}}\left[\mathcal{W}_2^2(\mathsf{proj}_v\,\mu, \mathsf{proj}_v\,\nu)\right] = \mathsf{SW}_2(\mu,\nu)
\end{aligned}$$

Let $\{x_i\}_{i=1}^\infty, \{y_j\}_{j=1}^\infty$ be infinite sets of points sampled from μ,ν, resp. Under the assumption that μ,ν have compact and bounded support, SW_2 is bounded above by the diameter of the support. Then, using the result above, we can apply Lebesgue's dominated convergence theorem and continuity of $\exp(\cdot)$ to interchange the limit and expectation:

$$\begin{aligned}
&\overline{\mathcal{K}}_\gamma^N(\mu,\nu) = \mathbb{E}_{\substack{x_i\sim\mu\\ y_j\sim\nu}}\left[\mathcal{K}_\gamma(\mu_N,\nu_N)\right] \\
&= \mathbb{E}_{\substack{x_i\sim\mu\\ y_j\sim\nu}}\left[e^{-\gamma \mathsf{SW}_2(\mu_N,\nu_N)}\right] \\
&\overset{\text{a.s.}}{\to} \mathbb{E}_{\mu,\nu}\left[e^{-\gamma \mathsf{SW}_2(\mu,\nu)}\right] = e^{-\gamma \mathsf{SW}_2(\mu,\nu)} = \mathcal{K}_\gamma(\mu,\nu).
\end{aligned}$$

5.2 Monte Carlo Wasserstein

While the sliced Wasserstein is very fast compared to the exact optimal transport distance, we will take the remainder of this chapter to attempt to find an even faster estimator for it. We will take inspiration from the idea of *multi-level Monte Carlo* methods.

5.2.1 Multi-level Monte Carlo

A classic technique in numerical simulations is the idea of an *control variate*. Assuming the goal is to compute the expectation value of a certain function P, one can achieve good empirical and theoretical performance by using a control variate Q, assuming $\mathsf{corr}(\mathsf{P}, \mathsf{Q})$ is large and $\mathbb{E}[\mathsf{Q}]$ is simple to compute [103]. The idea behind multi-level Monte Carlo algorithms is to leverage a coarse series of approximations to a given function, all of which are well-correlated with each other, to build an estimator that is more accurate than the sum of its part. More rigorously, consider some functional P, which is expensive to compute. If one can construct a hierarchy of functions $\mathsf{P}_\ell, \ell = 0, 1, 2, ...$ such that each subsequent function in the hierarchy has improved approximation accuracy but also increased cost, then the telescoping identity

$$\mathbb{E}[\mathsf{P}] = \mathbb{E}[\mathsf{P}_0] + \sum_{\ell=1}^{\infty} \mathbb{E}[\Delta\mathsf{P}_\ell],\ \Delta\mathsf{P}_\ell = \mathsf{P}_\ell - \mathsf{P}_{\ell-1}$$

trivially holds. Then, each expectation can be computed with an independent Monte Carlo integral, and the infinite sum can be truncated appropriately, giving the estimator:

$$\hat{\mathsf{P}} \approxeq \mathbb{E}[\mathsf{P}_0] + \sum_{\ell=1}^{\mathtt{L}} \frac{1}{\mathtt{n}_\ell} \sum_{\mathtt{i}=1}^{\mathtt{n}_\ell} (\Delta\mathsf{P}_\ell^{(\mathtt{i})}).$$

The computational savings that such a scheme may afford is given by the following

theorem, due to ref. [104].

Theorem 5.2.1 (Giles, 2013). *Denote by* $\mathsf{Y}_\ell = \frac{1}{n_\ell}\sum_{i=1}^{n_\ell}(\Delta \mathsf{P}_\ell^{(i)})$ *the Monte Carlo approximation to the hierarchical difference term. If the following conditions hold:*

1. $\|\mathbb{E}[\Delta \mathsf{Y}_\ell]\| \leqslant c_1 2^{-\alpha \ell}$
2. $\mathsf{Var}[\Delta \mathsf{Y}_\ell] \leqslant \frac{1}{N_\ell} c_2 2^{-\beta \ell}$
3. $\mathsf{cost}(\Delta \mathsf{Y}_\ell) \leqslant c_3 N_\ell 2^{\gamma \ell}$

such that $\alpha \geqslant \frac{1}{2}\min(\beta, \gamma)$*, then, there exists a constant* c_4 *such that the MLMC estimator achieves the error and cost bounds*

$$\mathbb{E}[(\hat{\mathsf{P}} - \mathbb{E}[\mathsf{P}])^2] \leqslant \epsilon^2 \text{ and } \mathsf{cost}(\hat{\mathsf{P}}) \leqslant \begin{cases} c_4 \epsilon^{-2} & \beta > \gamma \\ c_4 \epsilon^{-2} \log^2 \epsilon & \beta = \gamma \\ c_4 \epsilon^{-2-(\gamma-\beta)/\alpha} & \beta < \gamma \end{cases}$$

This suggests an efficient scheme for approximating the Wasserstein and sliced Wasserstein distances, when the number of points in the sample is very large or the measures are absolutely continuous. Specifically, let D_n, $n = 1, \ldots, \infty$ be the distance between the empirical distributions formed from n samples:

$$D_n := \mathsf{SW}_2\left(\frac{1}{n}\sum_{\ell=1}^{n}\delta_{x_\ell}, \frac{1}{n}\sum_{\ell=1}^{n}\delta_{y_\ell}\right).$$

In particular, we can write the telescoping sum:

$$\begin{aligned} \mathsf{SW}_2^2(\mu, \nu) &= \lim_{n\to\infty} \mathbb{E}[D_n^2] && (5.8) \\ &= \mathbb{E}[D_1^2] + \sum_{j=1}^{\infty}\left(\mathbb{E}[D_{n(j+1)}^2] - \mathbb{E}[D_{n(j)}^2]\right) && (5.9) \\ &= \mathbb{E}_{x,y}[D_1^2] + \mathbb{E}_{n\sim G}\mathbb{E}_{x,y}\left[\frac{1}{g_j}\left(D_{n(j+1)}^2 - D_{n(j)}^2\right)\right], && (5.10) \end{aligned}$$

where G is a chosen distribution in $\mathcal{P}_+(\mathbb{N})$ with density $g_1, g_2, \ldots$, and $n(j)$ is a monotonically increasing integer sequence with $n(1) = 1$.

Before we prove any results about this MLMC estimator, we note similar bounds present in the literature for the exact and entropic versions of the Wasserstein distance. In particular, known bounds on time complexity are shown in Table 5.1.

Method	Cost	Base metric
Empirical optimal transport [64]	$n^3\epsilon^{-2}$	Exact Wasserstein
Parallelizable optimal transport [105]	$n^2\epsilon^{-1}$	Exact Wasserstein
Sinkhorn iterations [66]	$n^2\epsilon^{-3}$	Entropic transport
Greenkhorn iterations [69]	$n^2\epsilon^{-2}$	Entropic transport
Projected dual mirror descent [106]	$n^{5/2}\epsilon^{-1}$	Entropic transport

Table 5.1: Bounds on time complexity for approximation schemes to standard and entropic Wasserstein.

To the best of our knowledge, there are no such results explicitly for the sliced Wasserstein distance.

Theorem 5.2.2. *The estimator defined in Eq. (5.8) is asymptotically unbiased and satisfies the following asymptotic relations:*

$$\mathbb{E}[\widehat{\mathsf{SW}} - \mathbb{E}[\mathsf{SW}]] \leqslant \epsilon^2 \implies \mathsf{cost}(\widehat{\mathsf{SW}}) \leqslant c\epsilon^{-3}$$

Proof. The asymptotic unbiasedness of the estimator follows directly from the fact that the Wasserstein distance metrizes weak convergence.

Proving the runtime bound is slightly more involved. For the remainder of this analysis, we will fix the number of projections to be some constant value, and ignore it in the asymptotic limit as the number of points $n \to \infty$. To prove the total time complexity, we require three scaling constants: the bias, the variance, and the cost.

Bias. In general dimension, the asymptotic scaling of the bias of Wasserstein distance is:

$$\|\mathcal{W}_p^p(\mu, \nu) - \mathcal{W}_p^p(\hat{\mu}_n, \hat{\nu}_n)\| = \mathcal{O}(n^{-1/d}).$$

In general dimension, this result has only recently been shown [65]. However, in one dimension, an analogous result dates back to 1969 [107]. It follows from the Glivenko-Cantelli convergence of the empirical quantiles to the true quantiles, which forms the basis for the closed form solution to the 1-d transport problem.

Variance. Here, we will specifically consider $\mathcal{W}_2$. To bound the variance, we will apply the Efron-Stein inequality. For generality, we

Theorem 5.2.3. *(Efron-Stein inequality.) Define* $X^n := (x_1, ..., x_n)$ *and* $\bar{X}_i^n := (x_1, ..., x_i', x_{i+1}, ..., x_n)$*, where* $x_i' \overset{\mathcal{D}}{=} x_i$ *are i.i.d. from the same distribution. For any function of a arbitrary vector-valued random variable* X^n*,*

$$\mathsf{Var}[f(X^n)] \leqslant \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\left[(f(X^n) - f(\bar{X}_i^n)^2\right]$$

We want a similar bound for $\mathsf{Var}[\mathcal{W}_2^2(X^n, Y^n)]$. As before, all the x_i are i.i.d. across i. By symmetry, we can set $i = 1$ without loss of generality and drop the index, writing $\bar{X}^n$ instead. The inequality then becomes:

$$\mathsf{Var}[\mathcal{W}_2^2(X^n, Y^n)] \leqslant \frac{n}{2} \cdot \mathbb{E}\left[(\mathcal{W}_2^2(X^n, Y^n) - \mathcal{W}_2^2(\bar{X}^n, \bar{Y}^n))^2\right]$$

We will move to the dual.

$$\begin{aligned} D_n^2(X^n, Y^n) - D_n^2(\bar{X}^n, \bar{Y}^n) &\leqslant \frac{1}{k}\left(\sup_{f,g\in\Phi} (f(x_1) + g(y_1)) - \sup_{f,g\in\Phi} (f(x_1') - g(y_1'))\right) \\ &\leqslant \frac{1}{k}(c(x_1, y_1) - c(x_1', y_1')) \end{aligned}$$

where we have used the definition of the constraint set Φ. Hence,

$$\begin{aligned} \mathbb{E}[\left(D_n^2(X^n, Y^n) - D_n(\bar{X}^n, \bar{Y}^n)\right)^2] &\leqslant \frac{1}{k^2} \cdot \mathbb{E}[(c(x_1, y_1) - c(x_1', y_1'))^2] \\ &\leqslant \frac{2}{k^2} \cdot \mathbb{E}_{\substack{x\sim\mu \\ y\sim\nu}}[c(x, y)^2] \end{aligned}$$

Assume that the cost $c(x, y)$ is the Euclidean distance. We will show that $\mathbb{E}[c(x, y)^2]$

is finite and independent of k.

Case 1. If μ has finite and bounded support, and the diameter of the support is bounded by R, then immediately $\mathbb{E}[c(x,y)^2] \leqslant R^2$.

Case 2. Instead, if μ has σ-subgaussian marginals, we first note the variance of the i-th dimension is bounded as $\mathbb{E}_{\mu_i}[(x^{(i)})^2] \leqslant \sigma^2$. In this case, we can write:

$$\begin{aligned}
\mathbb{E}[c(x,y)^2] &= \mathbb{E}\left[\sum_{i=1}^{d}(x^{(i)} - y^{(i)})^2\right] \\
&\leqslant \sum_{i=1}^{d}\left(\mathbb{E}[(x^{(i)})^2] + \mathbb{E}[(y^{(i)})^2] - 2\mathbb{E}[x^{(i)}y^{(i)}]\right) \\
&\leqslant \sum_{i=1}^{d}\left(\sigma^2 + \sigma^2 + 2\sigma\cdot\sigma\right) \\
&= 4d\sigma^2
\end{aligned}$$

where in the third line we have used the fact that μ, ν are σ-subgaussian in their marginals for the first two expectations, and for the third we have used Hölder's inequality. In either case, the expectation is finite and does not depend on k. Therefore, plugging back in, we have $\mathbb{E}[\left(D_n^2(X^n, Y^n) - D_n(\bar{X}^n, \bar{Y}^n)\right)^2] \leqslant \frac{2}{k^2}\cdot 4d\sigma^2$, which implies that:

$$\mathsf{Var}[D_n^2(X^n, Y^n)] \leqslant \frac{n}{2}\frac{2}{n^2}\cdot 4d\sigma^2 = \mathcal{O}(n^{-1})$$

which is the parametric rate[1].

Cost. As described above, the complexity of sliced transport is simply the same as that of sorting. Therefore, asymptotically, sliced transport costs $\mathcal{O}(n \log n)$ in the number of points.

These three results give us scaling constants $(\alpha, \beta, \gamma) = (1, 1, 2)$. Taking these scaling constants together, Theorem (5.2.1) implies that the estimator has the desired error and time complexity. □

[1] A similar argument due to ref. [65] can be applied to the primal problem to prove the same rate. And, while we do not reproduce it here, ref. [68] shows the same rate applies for the entropic transport problem as well.

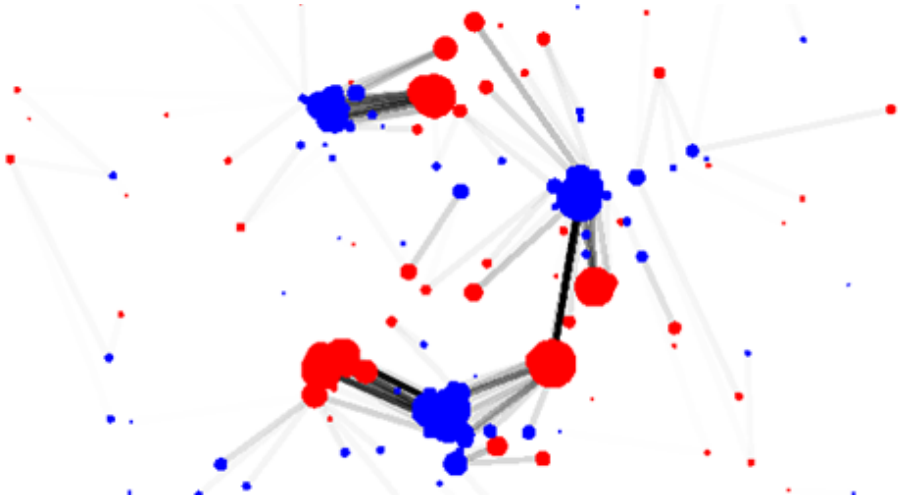
5.3 Conclusion

In summary, we have shown that there exists a kernel that satisfies the property of positive definiteness and therefore induces a feature space that is Euclidean. This kernel, based on the sliced Wasserstein distance, is well-suited to point clouds as it is well-behaved for empirical measures. Specifically, we prove that, for empirical approximations of absolutely continuous measures, the "bootstrapped" sliced Wasserstein kernel satisfies the Mercer condition to be a kernel. Further, it converges to the kernel similarity between the underlying measures in the limit as the number of samples goes to infinity. We additionally introduce a fast estimator based on the multi-level Monte Carlo method, and prove that its runtime is smaller than that of the sliced Wasserstein distance.

In the next section, we will test empirically the theoretical claims made in this section. We will show that these kernels are, in fact, useful in the specific application of anomaly detection.

Part III

Experimental results



The central goal of this work is to identify and detect anomalies. However, the definition of an anomaly itself is vague and varies from field to field. In particular, the machine learning scientist's vision of an anomaly is very different from that of a particle physicist. Anomalies at hadron colliders, as in the figure on the previous page (taken from ref. [108]) arise when a theoretical prediction is not realised in experimental data. In this sense, particle physicists are searching for anomalies in the aggregate – statistical deviations in a region of phase space from a background model. For example, regions of local overdensity are the telltale sign of a BSM particle, a phenomenon colloquially known as the "bump hunt." However, as we cannot directly recontruct the intermediate states of a resonant anomaly, there's no ground truth that assigns each event a label as anomalous or not, despite the best efforts of neural network-based taggers at the LHC to simulate this. In this part, we will dive into the distinction between these two frameworks, spending some time with the discriminative perspective on anomaly detection, before transitioning to the generative framework and mixture modelling.

This part will apply the framework that the previous chapters of this work have outlined to a series of anomaly detection tasks. Chapter 6 takes the transport-based kernels previously defined and leverages them in a pair of discriminative models. These models are then applied to detect cancerous cells in flow cytometry data, and to tag top quarks in simulated Pythia data. Chapter 7 is based on an unpublished work, and builds a generative model that uses the factorization theorem to statistically model dijet mixtures.

Chapter 6

Anomaly detection

> "We know that the only way to avoid error is to detect it and that the only way to detect it is to be free to inquire."
>
> —J. Robert Oppenheimer

In this section, we will apply our fast Wasserstein kernels to a pair of anomaly detection datasets Before we start, it is important to carefully define what we mean by an anomaly.[1] We will see that our definition has a huge impact on not only the types of models that demonstrate good performance, but also the kernels and techniques which we decide to use. To that end, we will distinguish between three types of anomalies.

1. **Localized density anomaly.** A localized abnormal density is a region of phase space that is either over- or under-dense relative to its surroundings.

2. **Outlier.** A spatial outlier is a point or group of points that is "far away" from the bulk of the background distribution, with respect to the base metric. This is distinguished from a localized abnormal density as a spatial outlier will lie in a region where density estimation is impractical, due to the lack of support.

3. **Something else entirely.** There are other types of anomalies as well – for example, a mixture model with one component having a small mixing fraction

[1] In keeping with the particle physics literature, we will sometimes refer to anomalies as the "signal".

could be considered an anomaly. However, it's very difficult to find these types of anomalies if we have no information about the background distribution, or some *a priori* guess about the shape or location of the signal component.

In the remainder of this chapter, we will numerically test out the kernels and approximations we have defined over the previous sections on a toy example of isotropic Gaussians. We will then introduce two real-world anomaly detection datasets – one based on cancer detection, and one from particle physics – and empirically test our kernels in two discriminative models. These models, the k-NN-LPE from [74] and the OC-SVM from [75], have shown good performance in a wide variety of anomaly detection tasks. However, as we will see, the kernel-based techniques in this section are best suited to address the first two types of anomalies.

6.1 The kernel zoo

In the previous chapter, we proposed several new Wasserstein-type distances and kernels. Before we dive into the task of anomaly detection, we first devote a brief moment to summarizing all the available kernels that operate on distributions, and their theoretical time complexity. As before, consider a pair of distributions $\widehat{\mu}, \widehat{\nu}$, each supported on n points in $\mathbb{R}^d$ (and for the bootstrapped kernels, $m \ll n$ subsampled points). For the sliced kernels, we consider t projections. In Table 6.1, we outline all the kernels considered in this analysis.

As a baseline, we have included the *kernel mean map*, proposed in ref. [109]. This kernel is defined as the map $\Phi : \mathcal{P}_+(\mathcal{X}) \to \mathcal{H}$ such that

$$\Phi(\rho) = \int_{\mathcal{X}} k(\mathbf{x}, \cdot) d\rho(\mathbf{x}).$$

where $\mathcal{H}$ is the RKHS of some kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. This suggests an efficient empirical method for computing the inner product in the feature space $\mathcal{H}$. In particular,

[2]Recall that the MLMC theorem 5.2.1 gives a computational cost in terms of the mean squared error of the approximation. Empirically, the runtime is upper bounded by the sliced Wasserstein complexity on the full sample.

Kernel	Time complexity	P.d.?	Unbiased?
Sliced Wasserstein (unbiased)	$\mathcal{O}(tn \log n)$	Yes	Yes
Sliced Wasserstein (biased)	$\mathcal{O}(tn \log n)$	Yes	No
Bootstrapped sliced Wasserstein	$\mathcal{O}(tm \log m)$	Yes	No
Bootstrapped coreset sliced Wasserstein	$\mathcal{O}(tm^2)$	Yes	No
Multilevel Monte Carlo sliced Wasserstein[2]	$\mathcal{O}(\epsilon^{-3}) \leqslant \mathcal{O}(tn \log n)$	Yes	Yes
Exact Wasserstein	$\mathcal{O}(n^3 \log n)$	No	Yes
Entropic Wasserstein (Sinkhorn divergence)	$\mathcal{O}(n^2)$	No	No
Kernel mean map	$\mathcal{O}(n^3)$	Yes	Yes

Table 6.1: A list of kernels operating on distributions and some of their properties, including if they are positive definite, if they are unbiased, and their theoretical time complexity.

$$\kappa(\mu, \nu) = \langle \Phi(\mu), \Phi(\nu) \rangle_{\mathcal{H}} = \frac{1}{n^2} \sum_i \sum_j k(x_i, y_j)$$

where $x \sim \mu, y \sim \nu$. This kernel is characteristic and universal, and when it is generated from the Gaussian kernel, the kernel mean map is equivalent to the moment generating function of a random variable drawn from the distribution μ [110]. Because of these properties, it has been repeatedly used as a kernel over the space of distributions in learning tasks [75].

6.1.1 Numerical experiments

In this section, we will study the empirical time complexity and numerical convergence of the kernels in our zoo. We will perform all our experiments on a toy dataset of isotropic Gaussians. This dataset is chosen as the Wasserstein, sliced Wasserstein, and kernel mean map all have easily computable closed form values [44]. The number of points in the dataset varies by experiment, but is bounded in the range $[1, 1 \times 10^6]$. All computations are performed on a 2016 MacBook Pro laptop, using commonly available implementations of the Wasserstein distance. In particular, the entropic (Sinkhorn) and exact Wasserstein kernels are implemented in the `PyOT` package [111] while the sliced Wasserstein kernel and kernel mean map are computed using built-in

functions in `SciPy` and `NumPy`

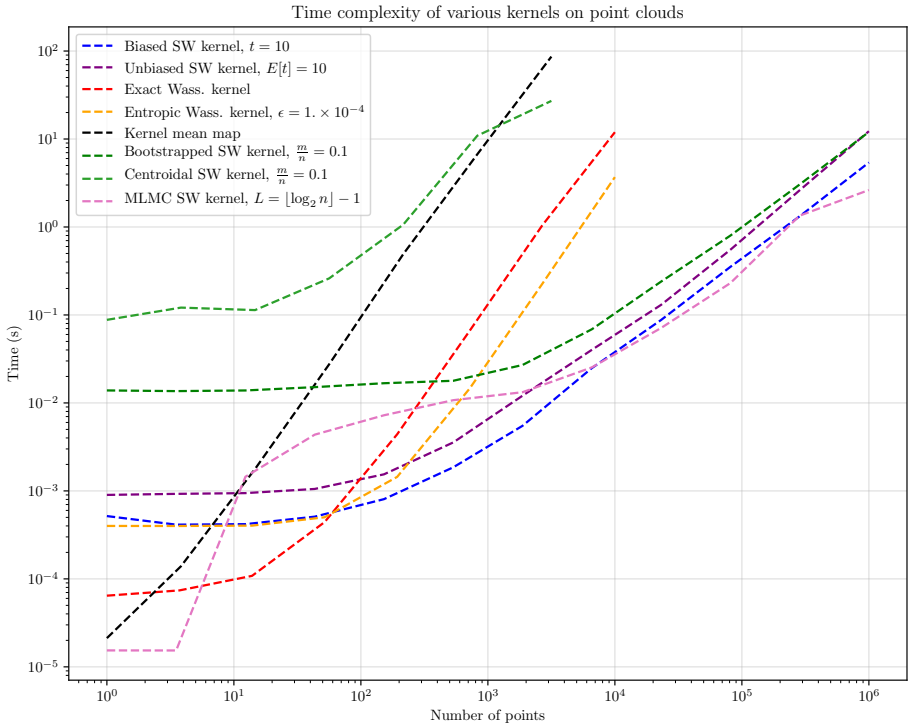


Figure 6-1: Wall time of various kernels computed between two isotropic Gaussians supported on varying numbers of points. Hyperparameters for each kernel were fixed prior to computation, to provide a fair comparison.

The plot in Fig. 6-1 shows the time complexity of running each kernel as a function of the number of points in the dataset. Immediately, we notice that the sliced Wasserstein kernels are empirically and theoretically the fastest kernels. The centroidal SW kernel, which is a version of the bootstrapped sliced Wasserstein kernel seeded through several iterations of Lloyd's algorithm, is the slowest, as the overhead of performing the k-means update steps dominates the actual transport computations. Behind that, the kernel mean map is the second slowest, and quickly outpaces it in the asymptotic limit. We remark that `PyOT` uses a highly optimized C implementation to solve the linear programs and compute the matrix balancing updates. As a result, for small n, these methods are among the fastest. However, they both share (approximately) the same asymptotic complexity as the kernel mean map, and as they are cubic, quickly become slower than the nearly-linear sliced distances. Finally, due to the high variance in the MLMC estimator in selecting the number of

levels in the hierarchy to traverse, we note unfortunately that it does not perform significantly better than the other sliced distances. A more careful analysis of the algorithm, akin to the variance reducing strategies employed for the unbiased SW kernel, may alleviate this issue.

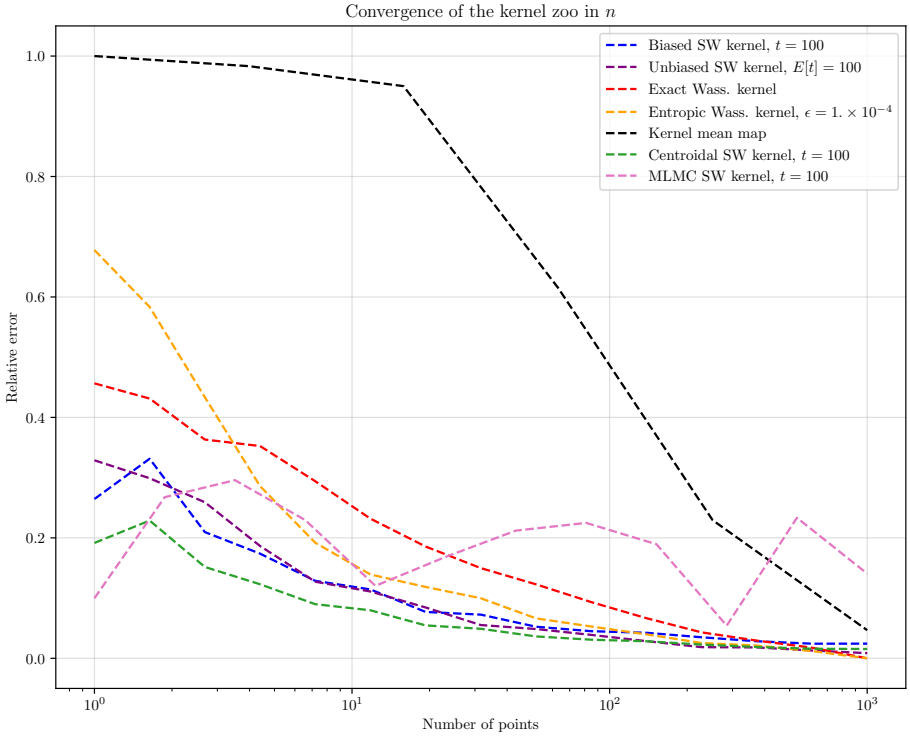


Figure 6-2: Convergence speed in n, the number of points in the empirical approximation, of the bootstrapped Wasserstein kernels relative to the true value. All kernels are computed between two isotropic Gaussians supported on varying numbers of points, with $\gamma = 0.1$. For the sliced Wasserstein distances, the number of slices is fixed (in expectation) at $t = 100$.

Next, we can observe the convergence speed of bootstrapped versions of these kernels with respect to the number of points subsampled. This is shown in Fig. 6-2. As we are considering isotropic Gaussians, the exact value of the Wasserstein distance can be computed from the Bures metric:

$$\mathcal{W}_2^2\left(\mathcal{N}(\mu_a, \Sigma_a), \mathcal{N}(\mu_b, \Sigma_b)\right) = \|\mu_a - \mu_b\|^2 + \left[\mathrm{tr}(\Sigma_a) + \mathrm{tr}(\Sigma_b) - 2\mathrm{tr}\left(\Sigma_a^{1/2}\Sigma_b\Sigma_a^{1/2}\right)^{1/2}\right]^{1/2}$$

and relative error is defined as the mean absolute deviation $\sum_i \frac{\|\hat{y}_i - y\|}{y}$. This is what we compare the exact Wasserstein distance to. The slowest to converge is the kernel

mean map, as it relies on an all-pairs computation. The fastest, as expected again, is the centroidal SW bootstrap. This follows directly from the theoretical arguments made in the previous chapter regarding the centroidal Voronoi tessellation and its relationship to Wasserstein barycenters. Therefore, it is expected that this will perform the best, compared to random samples. Unfortunately, the MLMC estimator again underperforms expectations. In the limit as N becomes large, the variance induced by the geometric distribution also becomes large, and the variance dominates the convergence. Finally, we note that all of these kernels have larger absolute error at smaller values of N. The unbiased kernel also has a nonzero error for small N, due to its increased variance. This is a consequence of the minimax argument stating that any empirical approximation of the Wasserstein distance converges in error with a $\mathcal{O}(\mathfrak{n}^{-1/\mathfrak{d}})$ rate. Chaining this with the triangle inequality gives a lower bound on the bias.

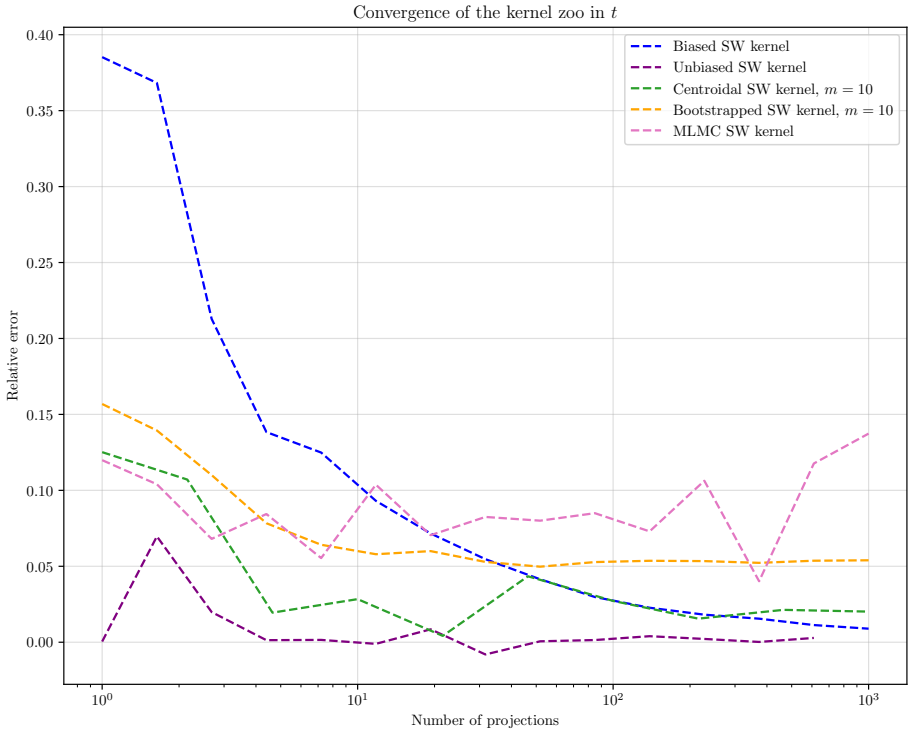


Figure 6-3: Convergence speed in $\mathfrak{t}$, the number of projections in the sliced Wasserstein kernels, relative to the true value. All kernels are computed between two isotropic Gaussians supported on $\mathfrak{n} = 50$ points, with $\gamma = 0.1$ fixed.

Finally, to conclude our discussion of the numerics, we consider the convergence of

the sliced kernels in terms of the number of slices, as shown in Fig 6-3. As expected, the unbiased sliced kernel converges immediately to the true kernel value, but exhibits high variance for all values of N. Further, the MLMC estimator again fails due to high variance. In addition, we note that the bootstrapped estimators are not asymptotically unbiased, if the number of points is held fixed at some finite value.

6.2 Datasets

In this section, we will briefly describe the two datasets that we consider for anomaly detection, and briefly provide representative visualizations of sample datapoints. While the first dataset is not related to physics, we include it as a demonstration of the different types of anomalies that one can encounter in point-cloud-based datasets (and, in particular, a real-world example that our anomaly detection techniques are well-suited to address). The second dataset is a standard tagging problem encountered at the Large Hadron Collider.

Flow cytometry. Flow cytometry is a method by which the properties of a suspension of particles in a fluid can be probed through optical scattering of a laser through individual particles, one at a time. It is commonly used in histology and oncology to understand cancerous tissues. For our purposes, the output of a flow cytometer can be thought of as a high-dimensional point cloud. Each element in the point cloud represents an individual cell, and each dimension represents the prevalence of a certain type of glycoprotein or immunoglobulin within the cell, as measure by the amount of light scattered at a certain (set of) wavelengths. In this analysis, we will use the datasets provided by ref. [112]. In this sample, 357 patients are tested for adult acute myeloid leukemia. For each patient, anywhere between 9,000 and 30,000 cells were collected, and intensity values were measured for 7 proteins of interest. We subsample the point cloud down so that each element of the dataset has 1,000 constituents. A sample 2-dimensional histogram for a pair of datapoints is presented in Fig. 6-4.

We note that the original dataset was intended for use in a supervised classification

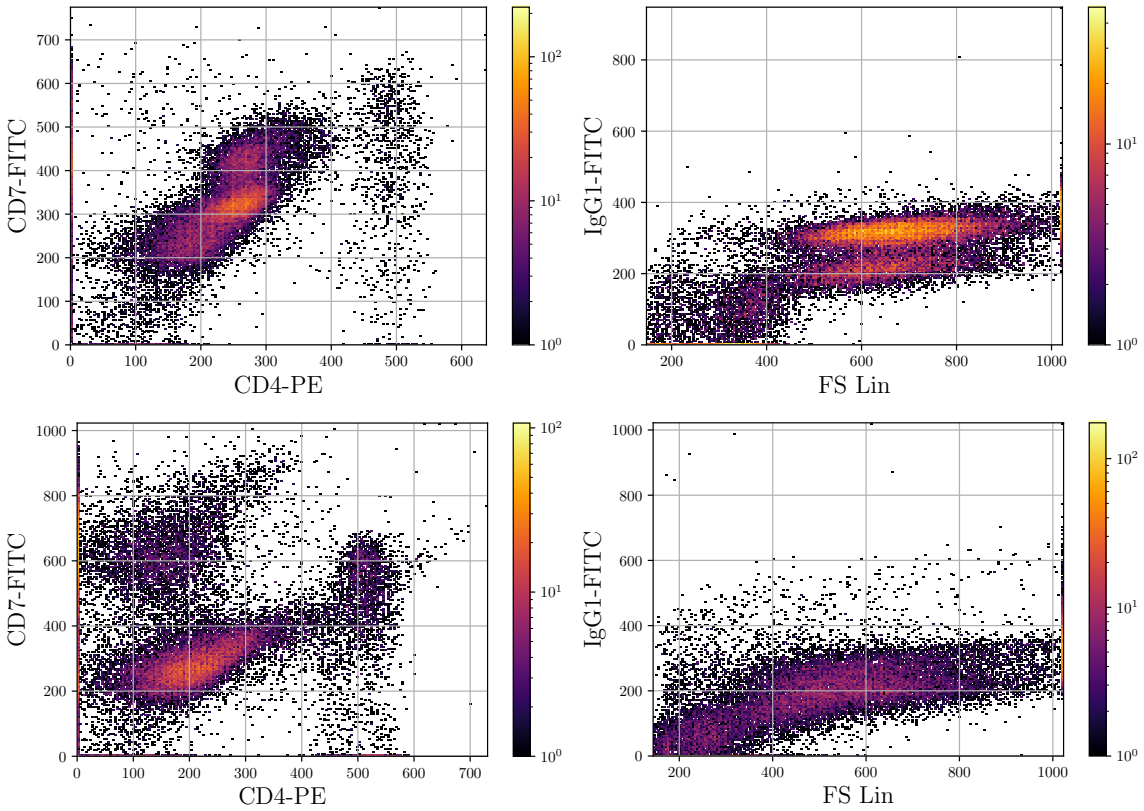


Figure 6-4: Two-dimensional marginals sliced through the flow cytometry data. The two histograms in the top row are taken from a patient suffering from AML, while the bottom row represents a normal sample.

study. For our purposes, as the fraction of AML-positive patients is $43/357 \approx 12.7\%$, we can consider the unsupervised anomaly detection problem instead. This problem is made more challenging by the small number of background samples, meaning density estimation will be difficult. However, as we will demonstrate later, we still achieve strong performance by a variety of metrics, although not comparable to the supervised methods.

Top quark tagging Our second dataset is a sample study of tagging top quarks against a QCD dijet background [113]. Top quarks, due to their relatively heavy mass, couple strongly to many theorized BSM particles. For example, the proposed *topcolor* theory suggests that a new force with $\mathsf{SU}(3)\otimes\mathsf{SU}(3)$ symmetry is spontaneously broken by a Z' boson, which decays to a top and anti-top quark [114]. Therefore, identifying

and tagging top quark jets can provide statistical evidence for this theory, and many others.

When dealing with point clouds in the jet plane, care must be taken to avoid symmetries. In fact, as the plane is invariant under reflections across both the $\mathfrak{y}$ and ϕ axes, as well as translations, we perform the following preprocessing steps. First, each jet is centered around the hardest particle in the event (i.e., the particle with the largest $\mathfrak{p}_T$). Next, the second hardest particle is placed in the upper left quadrant. Examples of these events are shown in Fig. 6-5.

This dataset contains 400,000 top quark jets and 400,000 QCD background jets. As with the flow cytometry data, this dataset was originally intended as a balanced supervised classification problem. However, by subsampling the signal class to be $\approx 10\%$ of the total dataset, we can induce an anomaly detection problem. Our final dataset is a random sampled set of 11,000 jets, 1,000 of which are top quarks. We can contrast this task with the flow cytometry dataset. We find reason to believe that top quark tagging will be more challenging for these models, for the following reasons

- **Large dataset.** The number of samples is an order of magnitude larger, and each individual sample has fewer elements (roughly 100). This meaning the computational tradeoff favors approximations of the kernel matrix instead of approximations to the kernel between any pair of data points, and our use of the sliced Wasserstein kernel is not as much of a computational advantage in this setting.

- **Entangled anomaly.** As evidenced by the plot in Fig. 6-5, standard observables like invariant jet mass have no discriminative power to tag individual top quarks against a dijet background without either a background density model or supervised training samples. As the anomaly is an overdensity embedded in the support of the background, it represents a fundamentally different type of anomaly than the flow cytometry data.

The discussion above suggests that anomaly detection techniques can be roughly broken down into two categories: discriminative and generative methods. The dis-

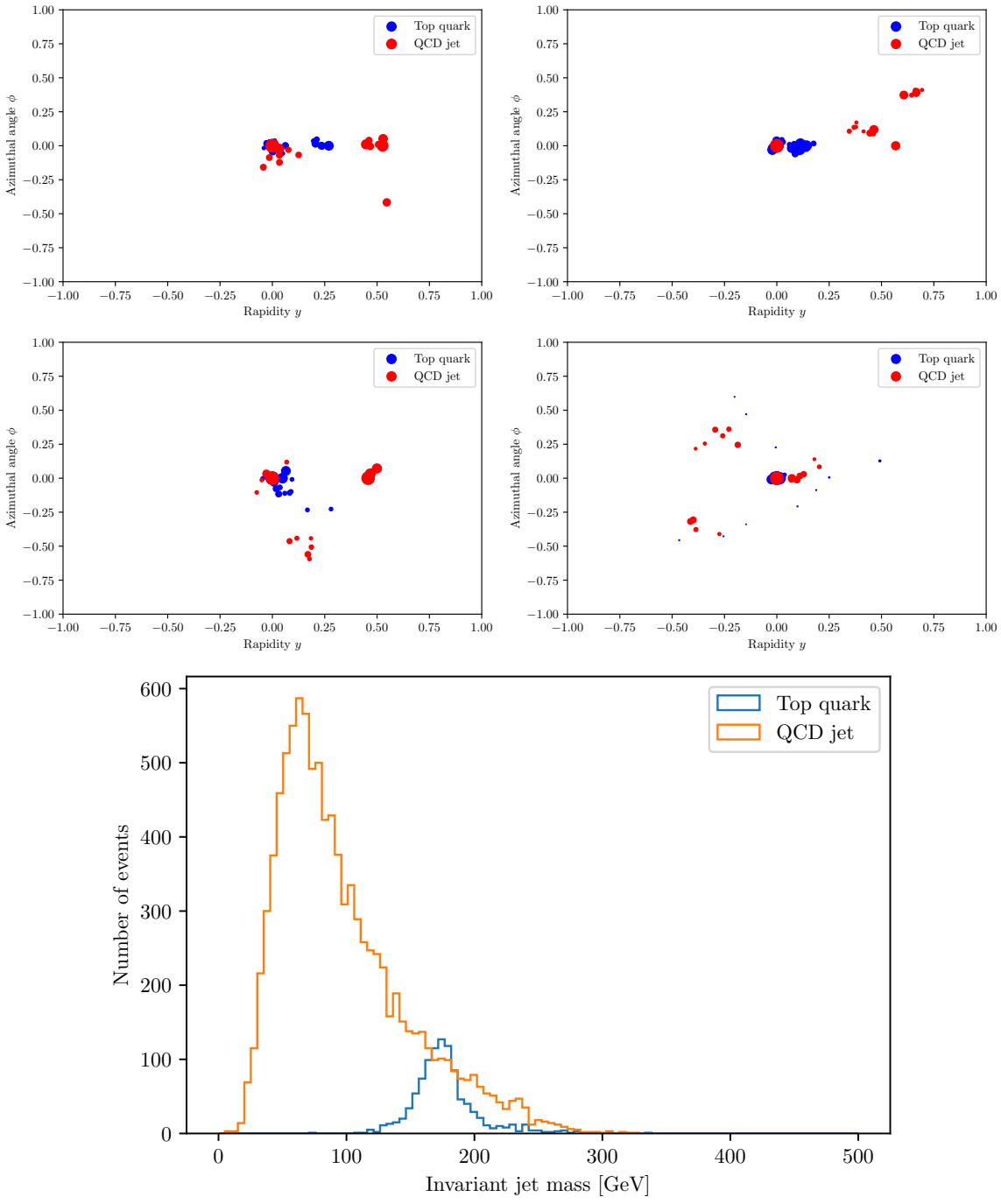


Figure 6-5: Four samples from the top tagging dataset, and the distributions of invariant jet mass, broken up into signal and background spectra.

tinction between the two can be roughly outlined as follows. In some cases, it is desirable to identify an individual event as anomalous. Under this assumption, we

seek to find the likelihood $\mathcal{L} = \frac{P(x|y=1)}{P(x|y=0)}$, where $y = 1$ denotes the signal class. This leads to a discriminative procedure, where we seek to classify individual instances as either anomalous or not. Contrast this to a scenario where each event is viewed as being generated by a mixture between two classes, one anomalous and one background. In this case, we are more interested in modeling the joint distributions between labels and observables $P(x, y)$. Instead of there being a ground truth for every event, in this paradigm we seek to bound the behavior of each generative process in a specific region of phase space. With this distinction established, we may proceed to applying these models to real-world datasets.

6.3 Discriminative methods

Equipped with a set of kernels, we can proceed to our main task. In this section, we will discuss two techniques for performing kernel-based anomaly detection through classifying samples, which constitute our discriminative models.

6.3.1 Nearest-neighbor density estimation

If the anomalies appear in regions of phase space where there are few surrounding samples, nearest-neighbor density estimation techniques are extremely powerful [74, 115, 116]. This family of techniques seeks to locally approximate the *outlier factor* through kernel-based density estimation. In particular, we will focus on the k-NN-LPE technique [74]. To perform this type of anomaly detection, we must first generate the kernelized k-NN graph, and assign the value $R(\mu_i) = \kappa(\mu_i, \mu_{i_k})$ as the distance to the k-th nearest neighbor for each event in the set. Then, each event receives a score defined as:

$$S(\mu_i) = \frac{1}{n} \sum_{\nu \neq \mu_i} \mathbb{1}\{R(\mu_i) \geqslant R(\nu_i)\}$$

.

The higher this score, the further away a point is from its k-nearest neighbors, and

therefore the more anomalous it is. Then, applying a threshold on these scores will give the most anomalous points. In previous work, the kernels used in this method have been Gaussian kernels based on the ℓ_2 and KL divergences, both of which fail to metrize weak convergence [75].

6.3.2 One-class support vector machines

Another interesting method for solving this problem is the method of one-class support vector machines (OC-SVM), first proposed by ref. [75].[3] The objective of the kernel-based one-class SVM is to find a maximum margin hyperplane in the feature space, $\mathbf{w} \in \mathcal{H}_\kappa$, that best separates the mapped data from the origin. This hyperplane is found by solving the SVM linear program:

$$\begin{aligned} \min_{\mathbf{w},\rho,\xi} \quad & \frac{1}{2}\langle \mathbf{w}, \mathbf{w}\rangle + \frac{1}{\nu n}\mathbb{1}^\intercal \xi - \rho \\ \text{s.t.} \quad & \langle \mathbf{w}, \Phi(\mu)\rangle \geqslant \rho - \xi \\ & \xi \geqslant 0 \end{aligned} \qquad \Longleftrightarrow \qquad \begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\boldsymbol{\alpha}^\intercal \mathbf{K} \boldsymbol{\alpha} \\ \text{s.t} \quad & 0 \leqslant \alpha_i \leqslant \frac{1}{\nu n} \\ & \mathbb{1}^\intercal \boldsymbol{\alpha} = 1. \end{aligned} \tag{6.1}$$

where ν is the fraction of anomalies expected, and n is the number of point clouds in the dataset. Here, as the feature mapping Φ is hard to compute, solving the problem in the dual space is preferred. It can be shown that this results in a nonlinear decision boundary encompassing the original data, and is closely related to the problem of finding a minimum enclosing sphere [75]. Then, to detect anomalies is as simple as checking on which side of the decision boundary the new data lies.

6.3.3 Results

We will first present results for the flow cytometry data, in Table 6.2. We note that, due to computational constraints, the number of subsampled points $m = 1000$ for all methods except for Centroidal SW and the kernel mean map, for which it is set at $m = 100$. Additionally, the number of projections is fixed using a hyperparameter

[3] When the kernel used is the kernel mean map, these are known as one-class support measure machines. We will use the terms interchangeably.

sweep at $\mathtt{t} = 50$, and the number of neighbors in the k-NN-LPE is set at $\mathtt{k} = 30$.

Flow cytometry						
OC-SVM				k-NN-LPE		
Kernel	Precision	Recall	F-1 score	Precision	Recall	F-1 score
Unbiased SW	0.67	0.67	0.67	**0.72**	**0.73**	**0.73**
Biased SW	0.66	0.66	0.66	0.70	0.73	0.72
Entropic OT	0.67	0.67	0.67	0.68	0.66	0.66
Exact OT	**0.67**	**0.68**	**0.67**	0.68	0.66	0.66
Centroidal SW	0.42	0.42	0.42	0.51	0.60	0.57
Kernel mean map	0.53	0.36	0.44	0.58	0.52	0.54
MLMC SW	0.28	0.22	0.26	0.18	0.18	0.18

Table 6.2: Results for discriminative methods on the flow cytometry dataset. The best performing algorithm is the k-NN-LPE, using the unbiased sliced Wasserstein kernel.

As evidenced in Table 6.2, these techniques show strong performance on the task of anomaly detection in flow cytometry data. As our primary metric, we use the concepts of precision, recall, and F-1 score, which are defined as follows:

$$\begin{aligned} \text{Precision} &= \frac{\text{True positives}}{\text{True positives} + \text{false positives}} \\ \text{Recall} &= \frac{\text{True positives}}{\text{True positives} + \text{false negatives}} \\ \text{F-1} &= \frac{1}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} \end{aligned}$$

While the supervised methods were able to classify the data perfectly, the F-1 scores reported in show that these models are well suited to the flow cytometry data. The performance of the unbiased sliced Wasserstein kernel is the best by a small but not insignificant margin, followed closely by the biased SW, and entropic and exact OT kernels. The other three kernels perform significantly worse, either due to their high variance or computational complexity. Between the k-NN-LPE and the OC-SVM, we find that the nearest neighbor density method performs slightly better. This suggests

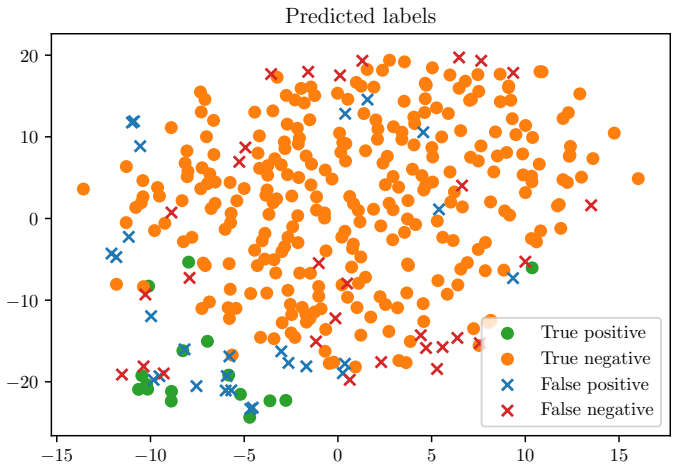


Figure 6-6: A t-SNE embedding of the flow cytometry dataset using the best-performing method, the k-NN-LPE using the unbiased sliced Wasserstein kernel. Datapoints are marked with the classification made by the method, and whether this was a correct prediction or not.

that the anomalous events are better described as spatially localized in some region of phase space, not isolated events far from the bulk of the background distribution. This hypothesis is supported by the t-SNE embedding presented in Fig. 6-6. t-SNE is a technique used to embed and visualize arbitrary manifolds in Euclidean space, while preserving clusters [117]. As shown, the anomalies are localized in the same region of the t-SNE plot.

Top quark tagging						
OC-SVM				k-NN-LPE		
Kernel	Precision	Recall	F-1 score	Precision	Recall	F-1 score
Unbiased SW	0.25	0.27	0.26	0.28	0.33	0.30
Exact OT	0.27	0.33	0.32	0.35	0.35	0.35
Exact OT (with Nystrom)	0.25	0.33	0.31	0.21	0.21	0.21

Table 6.3: Results for discriminative methods on the top quark tagging dataset. While the algorithms are statistically better than random, they do not perform well at the anomaly detection task.

However, we note that neither discriminative model is well suited for the top quark tagging task. As shown in Table 6.3, regardless of kernel, the F-1 scores are extremely low. This is caused by multiple factors, but most importantly, we find the top quark

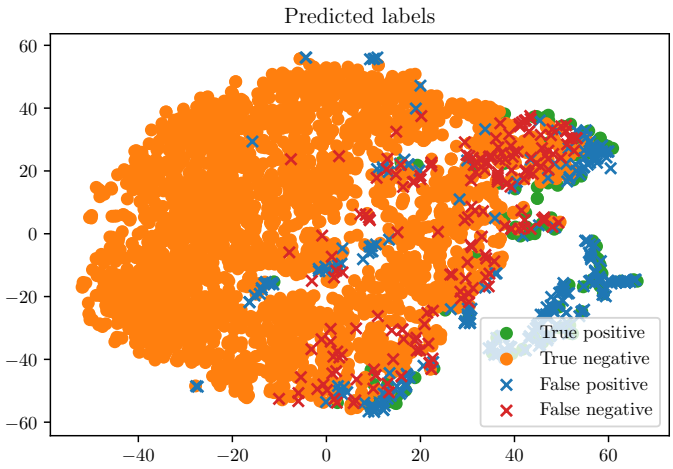


Figure 6-7: A t-SNE embedding of the top tagging dataset using the best-performing method, the k-NN-LPE using the exact Wasserstein kernel. Datapoints are marked with the classification made by the method, and whether this was a correct prediction or not.

anomaly is not an outlier in the traditional sense. In particular, as shown in Fig. 6-5, when the distribution is plotted for the invariant jet mass, the anomalous region has overlapping support with the background. In this projection, some background events are indistinguishable from the signal. Therefore, it is neither a localized anomalous density, nor is it a set of events far away in feature space – it falls into the third type of anomaly. This is the motivating reason for the "bump hunt" described in Chapter 2 – in a physically relevant observable, the anomaly is evidenced as an overdensity in a specific localised region of phase space. Our hypothesis was that, in the feature space induced by our set of kernels, we could reproduce this behavior. However, as clear from the t-SNE embeddings presented in Fig. 6-7, this is not the case. In fact, this means that despite the success of supervised nearest-neighbor tagging with the Wasserstein distance [108], searching for regions of anomalous density (as in the k-NN-LPE) or regions separated from the background support (as in the OCSVM) are not suitable for this specific anomaly detection task.

6.4 Conclusions

This section was primarily devoted to examining empirically the theoretical guarantees we established in previous sections for our set of kernels. We showed numerical experiments demonstrating the speed of convergence and time complexity of the kernels in our "kernel zoo." Then, we applied the kernels to a set of anomaly detection problems from disparate domains. We demonstrated three techniques, two discriminative and one generative, that each target a different type of anomaly. We claimed that the flow cytometry data could be well described as a localized density anomaly, and then experimentally verified this by showing that the nearest-neighbor-based anomaly detection system gave the best results for that dataset. In addition, we have demonstrated empirically that the sliced Wasserstein kernels are a strong and fast replacement to kernels based on exact and entropic transport distances.

Finally, we showed that tagging top quarks does not fall into the framework of either of these anomaly detection mechanisms. As the set of top quarks is neither localised in Wasserstein space nor spatially separated from the QCD dijet background, neither of our techniques demonstrates strong performance in detecting these anomalies. Therefore, we must find another model to separate this type of anomaly. In particular, as the support of the anomalous distribution overlaps so widely with the support of the background distribution, we should look towards a statistical picture of anomalies. In the next section, we will take this idea of generative modeling even further, and explicitly construct a topic model that leverages the factorization theorem to find anomalies.

Chapter 7

Factorized disentangling

> "From these considerations, it might seem that factorization should be easy to prove. It's not quite this simple, however..."
>
> —George Sterman

Before we begin this section, we will briefly recap the setting for our anomaly detection problem, when it comes to particle physics. Recall that we are operating on event signatures, comprised of jets. So far, we have considered classifying each jet individually as either anomalous or not. However, jets are fundamentally statistical in nature. It is difficult to isolate pure samples of a given type of jet from experimental data [118], and therefore predictions and analyses are often carried out under the assumption that collections of jets are a mixed sample from different types. Naturally, it is desirable to access pure samples from these mixed samples to understand the physics of the underlying particles. This idea, of *disentangling* samples, is the target of this section. This chapter is based on a joint work with Jesse Thaler, Eric Metodiev, and Patrick Komiske.

7.1 Introduction

We will focus on two specific tasks: quark/gluon discrimination, and resonant anomaly detection, both of which can be cast as disentangling problems.

- **Quarks and gluons** are the underlying particles that fragment into most jets

observed at the LHC. Jets produced from quarks and gluons are ubiquitious in both signal processes and the background QCD fragmentation. Unfortunately, there is no theoretical distinction between a "quark" or "gluon" jet. In practice, ground truths are often derived from parton shower event generators, but these are unphysical, and provide a barrier in comparing theory and experiment. For example, a recently proposed hadronic-level definition relies on the likelihood ratio between two mixed samples of quark and gluon jets [119], while models based on likelihood ratios have performed competitively across a variety of discrimination and tagging tasks [120, 121]. However, as data in practice is mixed, to model and understand quarks and gluons requires disentangling their component distributions from each other.

- **Resonant anomaly detection** is a fundamental tool in probing physics beyond the standard model. As jets are highly collimated, novel particles can be found as resonant "bumps" when plotted over the smoothly falling background of a specific observable. Machine learning provides an extremely useful tool in both searching for and validating the presence of a hypothesized anomaly [122–126]. Estimating the background density of a sample is crucial in this task to determine the significance against the null hypothesis. In this way, anomaly detection can be viewed as the task of disentangling the signal (if it exists) from the background.

In both of these problems, the underlying goal is to understand the connection between the parton or boson generated in a collision and the final-state jet it produces. This understanding can be leveraged to either tag and classify jets, or to make theoretical predictions about the properties of an observable conditioned on jet type. Fundamentally, both tasks fit well into the framework of *generative modeling*. In this paradigm, a mixed sample of jets $\mathcal{M}$ has a probability distribution over some observables $\mathbf{x}$ described as:

$$\mathsf{p}_{\mathcal{M}}(\mathbf{x}) = \sum_{\mathsf{k}} \mathsf{f}_{\mathcal{M}}(\mathsf{k}) \cdot \mathsf{p}_{\mathsf{k}}(\mathbf{x}) \tag{7.1}$$

where k is a jet type, and $\mathsf{f}(\mathsf{k})$ is its relative fraction. Therefore, the goal of both tasks is to recover the *components* p_k and the *mixing fractions* $\mathsf{f}(\mathsf{k})$. In recent years, supervised learning has provided a solution to this problem, when a sample of true (observable, label) pairs are known [113, 127–130], and many such methods have been implemented in practice at the LHC as taggers [131, 132]. These methods are *discriminative* in nature, like those in the previous section. Recall that this means that they seek to model the conditional distributions $\mathsf{p}(\mathbf{x}|\mathsf{k})$ directly. However, in both quark-gluon discrimination and anomaly detection, the ground truth labels are not known *a priori*, making this type of analysis difficult. Among unsupervised methods, *generative modeling* is a commonly used tool that provides a different approach. This framework seeks to model the joint distribution over observables and labels $\mathsf{p}_\mathsf{k}(\mathbf{x})$ instead. In particular, instead of labelling each individual jet with a category, generative modeling aims to directly understand the statistical distribution governing collections of jets of a certain type or types.

In this section, we propose and evaluate a new technique to statistically model and discriminate between dijets. Most LHC events are dijets, and understanding their behavior specifically is useful in both BSM searches [133, 134] and precision measurements [135]. Dijets are especially interesting as they satisfy a factorization theorem for jet substructure. Simply put, factorization implies that both the jets in a dijet event are statistically independent, conditioned on their joint types. For example, the value of the leading jet mass does not affect the subleading jet multiplicity, except through their coupling in type. This is an extremely powerful statement, motivated from first principles, which we will convert into a statistical constraint on a generative model. Starting from the generative model of "jet topics" [136], we leverage a factorization theorem for jet substructure to build a generative model and demonstrate a procedure for optimizing it to recover both the relative fractions of different types of jets within a mixed sample, as well as the component distribution for a given observable.

7.2 Factorized Topic Modeling

7.2.1 A review of factorization

To describe a generative model for dijet production, we will leverage *factorization*. Factorization, in words, is the statement that the cross section for dijet production can be decomposed as the product of independent probability functions. Each component of the cross-section corresponds to a different physical process contributing to the observed jet pair. Concretely, the cross-section can be written as follows [137]:

$$d\sigma = \sum_{ab\to cd} f_a(\xi_a) \otimes f_b(\xi_b) \otimes \mathcal{H} \otimes \mathcal{J}_c(z_c) \otimes \mathcal{J}_d(z_d), \tag{7.2}$$

where f are the standard parton distribution functions for the proton, ξ are the momentum fractions, $\mathcal{H}$ is the partonic cross section for the short-range hard scattering process ($ab \to cd$), and $\mathcal{J}$ are the jet branching functions. In this work, as our goal is jet tagging, we will primarily shift our focus to the part of this equation that governs jet substructure.

$$d\sigma \propto \sum_{c,d} \mathcal{H}_{c,d} \otimes \mathcal{J}_c(z_c) \otimes \mathcal{J}_d(z_d) \tag{7.3}$$

Our goal is to translate this physical theorem into a statistical constraint on the probability distribution over jet observables. For dijets, we will specifically consider each observation to be a pair $(\mathbf{x}_1, \mathbf{x_2})$, corresponding to the value of a given observable for the hardest and second-hardest jet in the event, respectively. Now, using equation (7.3) as a starting point, we will write down a generative model for dijet production.

7.2.2 A review of topic models

In this work, we will focus on *topic modeling*, a specific type of unsupervised mixture modeling. The goal of unsupervised mixture modeling is to identify the presence and characteristics of subpopulations found within a sample, without the presence of identifying labels for any individual datapoint within the sample. As we cannot

observe the intermediate partons that fragment to create jets, and it is challenging to isolate pure samples of any give type of jet [118], mixture modeling is a natural framework for analyzing events. To apply this framework in practice, we can utilize a powerful technique from the natural language processing community known as topic modeling. Topic modeling was first applied to jet physics in ref. [136]. Their work leveraged the statistical connection between themes in text corpora and jet flavors in event samples to propose a new data-driven method for defining classes of jets. In this section, we will briefly describe their setting, and some benefits of their framework. We will first consider an unfactorized topic model, in a single observable $\mathbf{x}$. For a mixed sample $\mathcal{M}$, this corresponds to a generative process with the following structure.

$$\begin{aligned} p_{\mathcal{M}}(\mathbf{x}) &= \sum_k f_{\mathcal{M}}(k) \cdot p_k(\mathbf{x}) \\ \text{s.t.} \quad \int_{\mathcal{X}} d\mathbf{x} p_k(\mathbf{x}) &= 1 \quad \forall k \\ \sum_k f_{\mathcal{M}}(k) &= 1 \end{aligned} \tag{7.4}$$

In the present setting of jet physics, each component k corresponds to a jet class (i.e., quark or gluon). The mixture components $\{p_k\}$ correspond to the distributions of any given jet observable $\mathbf{x}$, while the fractions $f(k)$ represent the fraction of the total sample which belongs to each component. The goal of a topic model is to simultaneously learn the components $\{p_k\}$ and fractions $f(k)$ from a set of samples $\{\mathcal{M}_i\}$.

Once the components and fractions are extracted, they can be used to construct an operational definition of jet classes, that only relies on cross-sectional data [119]. The optimal discriminant between two jet classes is given by the Neyman-Pearson lemma as:

$$L_{i/j}(\mathbf{x}) = \frac{p_i(\mathbf{x})}{p_j(\mathbf{x})} \tag{7.5}$$

Then, if two samples are mixtures of only two classes i, j, the likelihood ratio between them can be written simply as monotonic rescaling of $L_{i/j}(\mathbf{x})$. Equipped with this insight, the extrema of this likelihood ratio can be used to identify regions of phase

space that are enriched in a certain class. These enriched regions constitute the operational definition of those classes; in particular, the resultant classes are *mutually irreducible* distributions over the observable $\mathbf{x}$. Conveniently, the learned components from a topic model are, up to rescaling, exactly these classes. Therefore, topic modeling provides an effective and practical implementation of the operational definition. The theoretical guarantees that this framework enjoys also translate to experimental benefits, as well. As an example, topic modeling has been applied to the task of understanding quark and gluon jets at the LHC [138]. By using topic modeling, they were able to recover operationally-defined quark and gluon samples, and compare these classes to "quarks" and "gluons" tagged by a parton shower simulation. Through this comparison, regions of phase space where the simulation diverged from the observed data were identified.

7.2.3 Statistical considerations

In this section, we will build up a statistical formulation of our topic model. Unlike the univariate topic model described in eq. (7.4), we will operate on pairs of observables $\mathbf{x}_1, \mathbf{x}_2$, corresponding to the leading and subleading jets in an event. To begin, we note that a topic model is uniquely specified by a constrained generative process. The goal of a topic model is to learn a universal set of components $\mathsf{p}_\mathsf{k}(\mathbf{x}_1, \mathbf{x_2})$, and a sample-dependent set of mixing fractions $\mathsf{f}_\mathcal{M}(\mathsf{k})$, that accurately describe the data distribution of a given sample $\mathsf{p}_\mathcal{M}(\mathbf{x}_1, \mathbf{x}_2)$. To generate datapoints, then, the following procedure is followed:

1. Sample a category $\mathsf{k} \sim \mathsf{Multinomial}[\mathsf{f}_\mathcal{M}(1), ..., \mathsf{f}_\mathcal{M}(\mathsf{k})]$.

2. Sample a datapoint $(\mathbf{x}_1, \mathbf{x}_2) \sim \mathsf{p}_\mathsf{k}$.

This process yields the analogous formula for the sample distribution:

$$\mathsf{p}_\mathcal{M}(\mathbf{x}_1, \mathbf{x}_2) = \sum_\mathsf{k} \mathsf{f}_\mathcal{M}(\mathsf{k}) \mathsf{p}_\mathsf{k}(\mathbf{x}_1, \mathbf{x}_2). \tag{7.6}$$

To specify the form for $p(\mathbf{x}_1, \mathbf{x}_2)$, we must explicitly write down our constraints, which are as follows, in statistical language:

1. *Sample independence*: The model assumes that, to leading order, the jet observable $\mathbf{x}$ depends only on the initiating parton. We note that, in fact, there is some dependence on the process in addition to the flavor. As an example, in the case of a Z+jet emission, soft gluon resummation subtly changes the p_T spectrum of the light quark jet relative to the QCD background. However, experimental studies have shown a high degree of empirical independence, and we suggest that these differences can be considered negligible for our model [119]. For the case of QCD dijets, sample independence tells us that for any given jet, its distribution is a function of the initiation parton (either light quark or gluon) and its momentum fraction. Rigorously, if we define $p^{(1)}, p^{(2)}$ the distribution functions for the hardest and second-hardest jet, respectively, then the statement above can be written as:

$$p_k^{(1)}(\mathbf{x}; \xi) = p_k^{(2)}(\mathbf{x}; \xi). \tag{7.7}$$

2. *Factorization* tells us that the two jets in an event are statistically independent, conditioned on convolution through the matrix element describing the short-range scattering. From a statistical perspective, the factorization theorem given above is mathematically equivalent to stating that our topic model for dijets must be an *mixture of products*. Hence, this can be written as:

$$(\mathbf{x}_1 | k_1, k_2) \perp (\mathbf{x}_2 | k_1, k_2) \implies p(\mathbf{x}_1, \mathbf{x}_2) \propto \sum_{k_1, k_2} f(k_1, k_2) \cdot p_{k_1}^{(1)}(\mathbf{x}_1) \cdot p_{k_2}^{(2)}(\mathbf{x}_2). \tag{7.8}$$

Note that by simply replacing the structure of the sample-level probability distribution in Problem 7.6 with the constraints from Section 7.2, the mapping between the factorization theorem and statistical language can directly give us a topic model.

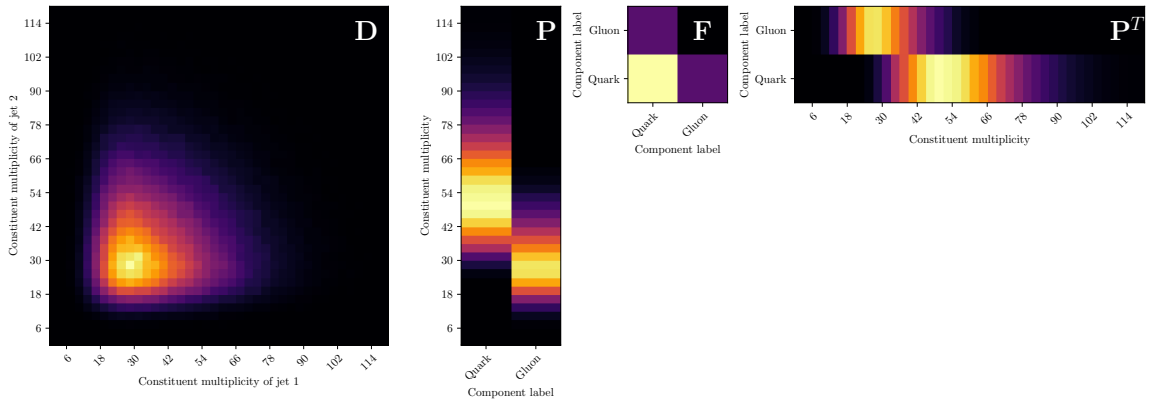


Figure 7-1: The paired observables from a dijet sample can be represented as a histogram, shown as the matrix $\mathbf{D}$. The generative process we describe can be visualized as the matrix product $\mathbf{PFP}^{\mathsf{T}}$, shown as a decomposition on the right.

The model can be expressed as follows.

$$\mathsf{p}_{\mathcal{M}}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{\mathsf{k}_1, \mathsf{k}_2} \mathsf{f}_{\mathcal{M}}(\mathsf{k}_1, \mathsf{k}_2) \cdot \mathsf{p}_{\mathsf{k}_1}(\mathbf{x}_1) \cdot \mathsf{p}_{\mathsf{k}_2}(\mathbf{x}_2). \tag{7.9}$$

However, the model itself is simply an encoding of the constraints. Our goal is to find the parameters of the model that give the best fit to the true distribution for the mixed sample $\mathsf{p}_{\mathcal{M}}$. Hence, the problem we seek to solve can be written as:

$$\begin{aligned} \min_{\mathsf{f}_{\mathcal{M}}, \{\mathsf{p}_{\mathsf{k}}\}} \quad & \mathsf{dist}\left(\mathsf{p}_{\mathcal{M}}(\mathbf{x}_1, \mathbf{x}_2) \middle\| \sum_{\mathsf{k}_1, \mathsf{k}_2} \mathsf{f}_{\mathcal{M}}(\mathsf{k}_1, \mathsf{k}_2) \cdot \mathsf{p}_{\mathsf{k}_1}(\mathbf{x}_1) \cdot \mathsf{p}_{\mathsf{k}_2}(\mathbf{x}_2)\right) \\ \text{s.t.} \quad & \int_{\mathcal{X}} \mathbf{dx} \mathsf{p}_{\mathsf{k}}(\mathbf{x}) = 1 \quad \forall \mathsf{k} \\ & \sum_{\mathsf{k}_1, \mathsf{k}_2} \mathsf{f}(\mathsf{k}_1, \mathsf{k}_2) = 1 \end{aligned} \tag{7.10}$$

for some appropriate distance between measures.

7.2.4 Disentangling topics with histograms

The model described in equation (7.4) suffers from a problem known as *non-identifiability*, which limits our ability to recover the true components and mixing fractions given a sample of data. In this section, we will define an identifiable model as well as the

assumptions necessary to avoid model degeneracy.

Identifiability. Consider the (infinite-dimensional) parameter space for our topic model. It is defined as

$$\Theta = \left\{ \left(\mathsf{f}, \{\mathsf{p}_\mathsf{k}\}\right) \middle| \int_\mathcal{X} \mathrm{d}\mathbf{x} \mathsf{p}_\mathsf{k}(\mathbf{x}) = 1, \sum_\mathsf{k} \mathsf{f}(\mathsf{k}) = 1 \right\}. \tag{7.11}$$

In the limit that we are given an infinite number of samples from the true distribution $\mathsf{p}_\mathcal{M}$, there may exist multiple parameters $\theta_\mathsf{i} \in \Theta$ for which the corresponding models yield an equivalent fit to the data. Each of these solutions would necessarily be interchangeable by some transformation of the mixing weights or components. A model that has this property is non-identifiable. We will outline three ways that our topic model, as specified above, can be non-identifiable.

Mutual reducibility. If one mixture component can be written as a linear combination of the remaining components, then the model is not uniquely identifiable, as there is a equivalent family of solutions that are equivalent up to reweighting. In particular, let $\mathsf{p}_\mathsf{k}(\mathbf{x}) = \sum_\mathsf{i} \alpha_\mathsf{i} \mathsf{p}_\mathsf{i}(\mathbf{x})$. Then, we can write:

$$\mathsf{p}_\mathcal{M}(\mathbf{x}) = \mathsf{f}_\mathcal{M}(\mathsf{k})\mathsf{p}_\mathsf{k}(\mathbf{x}) + \sum_\mathsf{i} \mathsf{f}_\mathcal{M}(\mathsf{i})\mathsf{p}_\mathsf{i}(\mathbf{x}) \tag{7.12}$$

$$= \sum_\mathsf{i} \left(\alpha_\mathsf{i} \mathsf{f}_\mathcal{M}(\mathsf{k}) + \mathsf{f}_\mathcal{M}(\mathsf{i}) \right) \mathsf{p}_\mathsf{i}(\mathbf{x}) \tag{7.13}$$

Hence, there are multiple equivalent ways of specifying the same model.

Degenerate mixing fractions. If two components share the same mixing coefficient, i.e., $\mathsf{f}_\mathcal{M}(\mathsf{k}_1) = \mathsf{f}_\mathcal{M}(\mathsf{k}_2)$, then the model is again not unique. In particular, we can write any linear combination of the two components:

$$\mathsf{p}'_{\mathsf{k}_1}(\mathbf{x}) = \alpha \mathsf{p}_{\mathsf{k}_1}(\mathbf{x}) + (1-\alpha)\mathsf{p}_{\mathsf{k}_2}(\mathbf{x}) \tag{7.14}$$

$$\mathsf{p}'_{\mathsf{k}_2}(\mathbf{x}) = (1-\alpha)\mathsf{p}_{\mathsf{k}_1}(\mathbf{x}) + \alpha \mathsf{p}_{\mathsf{k}_2}(\mathbf{x}) \tag{7.15}$$

that yields the same sample-level probability distribution.

For the remainder of this work, we will assume that both of the conditions above do not hold for the data we are presented. Specifically, we assume that the true generative process satisfies $f(k_1) \neq f(k_2)$, for all $k_1 \neq k_2$. Mutual reducibility is slightly harder to avoid. For the unfactored topic model described above and in ref. [136], it can be shown that the existence of a region of phase space where each component is uniquely supported suffices to make the components mutually irreducible. This is referred to in the literature as the *separability assumption* [139, 140]. For the purposes of this work, we will not assume that separability holds. When we discuss quark-gluon discrimination, we will show that the components are recoverable even without this assumption. In the case of a factorized topic model, one additional property of the true mixing fractions can lead to degeneracy.

Super-factorization If the true mixing matrix $\mathbf{F}$ satisfies the condition $\text{rank}(F) = t \leqslant k$, then there is any decomposition into t components that fits the observed data as well as any model with k components. We refer to this case as *"super-factorization."* In particular, if the mixing fractions can be written as

$$\begin{aligned} f_{\mathcal{M}}(q, q) &= a^2 \\ f_{\mathcal{M}}(q, g) &= a(1-a) \\ f_{\mathcal{M}}(g, g) &= (1-a)^2 \end{aligned} \tag{7.16}$$

for some fraction $a \in [0, 1]$, then quarks and gluons cannot be discriminated by our model, as the model is equivalent to the product of a single component $p_{\text{mixed}} = a p_q(\mathbf{x}) + (1-a) p_g(\mathbf{x})$ with itself.

This scenario is the factorized analog to non-separable component distributions. In practice, the minimization problem (7.10) as written is intractable, as the distribution $p_{\mathcal{M}}$ is continuous. As we only have access to it through a finite number of samples, and it is impossible to optimize over the space of all continuous probability distributions, we must find some way of operating on both $p_{\mathcal{M}}$ and the components p_k. There are two reasonable choices for this reformulation, outlined below.

1. *Parametric modeling.* In this paradigm, we force a distributional form on p_k, governed by some finite-dimensional parameter $\theta \in \Theta$. For example, one common mixture model can be defined as:

$$\begin{aligned} \min_{\theta_k \in \Theta, f} \quad & \mathsf{dist}\left(\mathsf{p}(\mathbf{x}), \sum_k \mathsf{f}(k)\mathsf{p}_k(\mathbf{x})\right) \\ \text{s.t.} \quad & \mathsf{p}_k(\mathbf{x}; \theta_k) = \mathcal{N}(\mathbf{x}; \mu_k, \sigma_k) \end{aligned}$$

   where the minimization is performed in parameter space. This method gives the well-known Gaussian mixture model, which is solvable by applying a projected descent algorithm. However, for our purposes, it is impractical to assume a parametric form for the shape of arbitrary jet observables. Therefore, we will not discuss this technique further in the present work.

2. *Non-parametric modeling.* One can instead discretize each distribution into a histogram. The advantages of this method are numerous. Primarily, it frees us from enforcing an implicit prior on the system in the form of a parametric assumption on distribution shape. It also allows us to improve computational performance, by reformulating the model as a matrix decomposition.

Hence, we prefer the second option, and move to the non-parametric setting. Define the matrix $\mathbf{D}$ to be the 2-dimensional histogram generated by jointly binning the sampled data across $\mathbf{x}_1$ and $\mathbf{x}_2$. Similarly, let $\mathbf{P}$ be the matrix whose columns are n-bin histograms representing each component $\mathbf{p}_k$. By rewriting the equation above in terms of histograms and bins, we arrive at the following non-convex program:

Problem 7.2.1.

$$\begin{aligned} \min_{\substack{\mathbf{F}\in\mathbb{R}^{k\times k}\\ \mathbf{P}\in\mathbb{R}^{n\times k}}} \quad & \|\mathbf{D}-\mathbf{P}\mathbf{F}\mathbf{P}^{\intercal}\|_{F}^{2} \\ \text{s.t.} \quad & \mathbf{P}^{\intercal}\mathbb{1}_{n}=\mathbb{1}_{k} \\ & \mathbb{1}_{k}^{\intercal}\mathbf{F}\mathbb{1}_{k}=1 \\ & \mathbf{P},\mathbf{F}\geqslant 0 \end{aligned}$$

where $\mathbb{1}_n$ is the n-dimensional vector of all ones and we have taken the Frobenius norm $\|\mathbf{A}-\mathbf{B}\|_F=\sqrt{\sum_{ij}(\mathbf{A}_{ij}-\mathbf{B}_{ij})^2}$ as our measure of distance. A pictorial representation of this discretization is given in Figure 7-1.

This problem is non-convex in $\mathbf{P}$ and $\mathbf{F}$, meaning finding global optima is not guaranteed. However, a rich variety of algorithms to solve the problem exists, most based on alternating minimization and gradient descent to quickly find local optima [141–148]. In the next section, we provide additional detail on the computational techniques used to solve this problem explicitly.

7.2.5 Algorithmic considerations

In this section, we will describe two methods to solve the topic modeling problem (7.2.1).

Symmetric tri-factorization. First, we consider the unconstrained alternating update rules due to ref. [144].

$$\begin{aligned} \mathbf{F}_{ij} &\leftarrow \mathbf{F}_{ij}\frac{(\mathbf{P}^{\intercal}\mathbf{H}\mathbf{P})_{ij}}{(\mathbf{P}^{\intercal}\mathbf{P}\mathbf{F}\mathbf{P}^{\intercal}\mathbf{P})_{ij}} \\ \mathbf{P}_{ij} &\leftarrow \mathbf{P}_{ij}\left(1-\beta+\beta\frac{(\mathbf{H}\mathbf{P}\mathbf{F})_{ij}}{(\mathbf{P}\mathbf{F}\mathbf{P}^{\intercal}\mathbf{P}\mathbf{F})_{ij}}\right) \end{aligned} \tag{7.17}$$

for some weight $\beta\in[0,1]$. To remain within the constraint set, we simply normalise $\mathbf{P},\mathbf{F}$ at the end of each update so that they satisfy the conditions given. This method will converge to a fixed point by the Karush-Kuhn-Tucker optimality conditions [149]. Further, the fixed point is guaranteed to be at least a local optimum.

Asymmetric NMF. Alternately, we explore a relaxation of the problem where the decomposition is not constrained to be symmetric. In particular, we solve the standard NMF problem 7.18 for two different matrices $\mathbf{C}, \mathbf{G}$. Problems of this form are studied in the signal processing community as *blind source separation* [150, 151].

$$\begin{aligned} \min_{\substack{\mathbf{C}\in\mathbb{R}^{N\times K}\\ \mathbf{G}\in\mathbb{R}^{N\times K}}} \quad & \|\mathbf{H}-\mathbf{C}\mathbf{G}^{\mathsf{T}}\|_{\mathsf{F}}^{2} \\ \text{s.t.} \quad & \mathbf{C}^{\mathsf{T}}\mathbb{1}_{n}=\mathbb{1}_{k} \\ & \mathbf{C},\mathbf{G}\geqslant 0 \end{aligned} \tag{7.18}$$

where ϵ is a small regularization parameter. We will now demonstrate an equivalence between optimal solutions to each of these problems.

Theorem 7.2.1. *Any optimal solution to Problem* (7.18) *with rank at least k can be transformed to an element of the constraint set of Problem* (7.2.1) *with rank k.*

Proof. First, we will establish that dropping the explicit symmetry constraint does not affect the symmetry of the optimal solution. Assuming the input matrix is positive semidefinite, any local search algorithm will still yield a symmetric solution while improving the rate of convergence [152]. In the more general case, we note the following proposition due to ref. [153].

Proposition 7.2.2. *Let the eigenvalues of* $\mathbf{H}$ *be* $\lambda_1, \lambda_2 ..., \lambda_n$. *Assume* $\lambda_i \neq -\lambda_j$ *for every nonzero* λ_i, λ_j. *Then the minimizing solution* $\hat{\mathbf{H}} = \mathbf{C}\mathbf{G}^{\mathsf{T}}$. *to Problem* (7.18) *is symmetric.*

Note that:

$$\mathbf{G}^{\mathsf{T}} = \mathbf{X}\mathbf{C}^{\mathsf{T}} \implies \mathbf{C}^{\dagger}\mathbf{G} = \mathbf{X} \tag{7.19}$$

which means that $\hat{\mathbf{H}} = \mathbf{C}\mathbf{X}\mathbf{C}^{\mathsf{T}}$. As $\mathbf{C}\mathbf{G}^{\mathsf{T}}$ is symmetric, therefore $\mathbf{X}$ is also symmetric. Then, performing the following rescaling:

$$\begin{aligned} \mathbf{P}_{ij} &\leftarrow \frac{\mathbf{C}_{ij}}{\sum_k \mathbf{C}_{kj}} \\ \mathbf{F}_{ij} &\leftarrow \mathbf{X}_{ij} \cdot \left(\sum_k \mathbf{C}_{ki} \sum_k \mathbf{C}_{kj} \right) \end{aligned}$$

we retrieve a pair $\mathbf{P}, \mathbf{F}$ that lies in the constraint set for Problem (7.2.1). $\square$

The result above suggests that the optimal solution to problem (7.18) is not necessarily optimal for problem (7.2.1). However, our reformulation is, in fact, empirically useful. Alternating non-negative least squares methods, like the two described above, are significantly less likely to approach global optima when the problem is non-convex (as in the latter) [154, 155]. Instead, we can leverage conic solvers designed for bi-convex problems [156]. This leads to both better theoretical convergence rates, as well as improved empirical performance. For this reason, we observe better empirical performance with the asymmetric formulation of the problem, and use that method throughout this work. As a final note, for the specific case of quark-gluon discrimination, we have exactly two components that we wish to learn. In the case that $k = 2$, there exist certain algorithmic techniques to improve the performance and increase the efficiency of these algorithms [157, 158]. Although we do not implement them, they are an interesting direction for future work.

Finally, we will address the separability assumption described in some more detail. Rigorously, a mixture component is t-separable if the following holds:

$$p_k \text{ is } t\text{-separable} \iff \int_R dp_k \geqslant t \text{ and } p_{k'}(R) = 0 \text{ for some region } R \subset \mathcal{X}. \quad (7.20)$$

However, recent work has shown that this assumption is not strictly necessary to retrieve the true solution, if it is unique. The "catch-words" algorithm due to ref. [159] relaxes separability, instead requiring only the existence of regions of phase space that are highly correlated with each other under a single topic. Similarly, in the case of finite mixture models, ref. [160] show that, as long as the size of each sample is super-exponential in the number of topics, the component distributions can be arbitrarily close. While we do not implement either of these models (their empirical complexity and runtimes are prohibitively large, and they are not easily generalized to the factorized setting), we mention them to show that it is possible, in theory, to relax the separability assumption.

In the case of (approximate) super-factorization, we note one final modification

necessary to solve the problem above. Unfortunately, due to the stochastic nature of the algorithm to solve Problem 7.2.1 (and the fact that finding a global optimum is not guaranteed) even if the mixing fractions are only approximately low-rank, finding a solution is empirically difficult. In this case, we increase the number of mixed samples to help the algorithm find the correct optimum, if it is unique. It suffices to bin on an auxiliary variable so that the mixing fractions in each bin are different. Let the mixed samples induced by these bins be $\{\mathcal{M}_i\}_{i=1}^B$. As established before, sample independence mandates that $\mathfrak{p}_k$ are independent of the mixed sample $\mathcal{M}_i$, while the mixing fractions are not. Thus, the binned version of the minimization problem can be formulated as:

$$\begin{aligned} \min_{\substack{\mathbf{F}_i \in \mathbb{R}^{k\times k} \\ \mathbf{P}\in\mathbb{R}^{n\times k}}} \quad & \sum_i \|\mathbf{D}_i - \mathbf{P}\mathbf{F}_i\mathbf{P}^\intercal\|_F^2 \\ \text{s.t.} \quad & \mathbf{P}^\intercal \mathbb{1}_n = \mathbb{1}_k \\ & \mathbb{1}_k^\intercal \mathbf{F}_i \mathbb{1}_k = 1 \\ & \mathbf{P}, \mathbf{F} \geqslant 0 \end{aligned} \tag{7.21}$$

where the mixing fractions are learned for each bin, but the components are learned jointly across all bins[1]. As an example, in the case of quark-gluon discrimination, it is well-known that forward jets are quark-enriched compared to central jets. Therefore, we can bin along quantiles in rapidity $\mathfrak{y}$. Empirically, we observe that the necessary number of bins to achieve good performance is small – anywhere between 3 and 7 bins is sufficient.

We now demonstrate the performance of this model on realistic quark and gluon samples.

[1]Similarly, one can apply the algorithm above to simultaneously learn multiple component distributions corresponding to different jet observables as well, noting that the mixing fractions are dependent only on the bin and independent of the observable. While this provides some additional discrimination power, it is not significant compared to the induced computational burden.

7.3 Quark and gluon disentangling

7.3.1 Event generation

The parton shower PYTHIA [161] is used to generate millions of dijet samples at $\sqrt{s} = 14$ TeV, including hadronization and multi-parton interactions. All detector-stable particles in the final state are clustered using the anti-k_t algorithm with $R = 0.4$ using FASTJET [162]. The two hardest jets in the event are selected if their total transverse momentum is within the range $p_T^{\text{sum}} \in [950, 1050]$ GeV and the ratio of their momenta $p_T^{(1)}/p_T^{(0)} > 0.8$. Additionally, we add a rapidity cut on the range $|y| \leqslant 2$. This yields a total of 1,432,784 total jet pairs, which constitute our mixed sample $\mathcal{M}$. Empirically, we observe that the true mixing fractions are

$$f_{\mathcal{M}}(g, g) \approx 0.13, \quad f_{\mathcal{M}}(q, q) \approx 0.41, \quad f_{\mathcal{M}}(q, g) = f_{\mathcal{M}}(g, q) \approx 0.23. \tag{7.22}$$

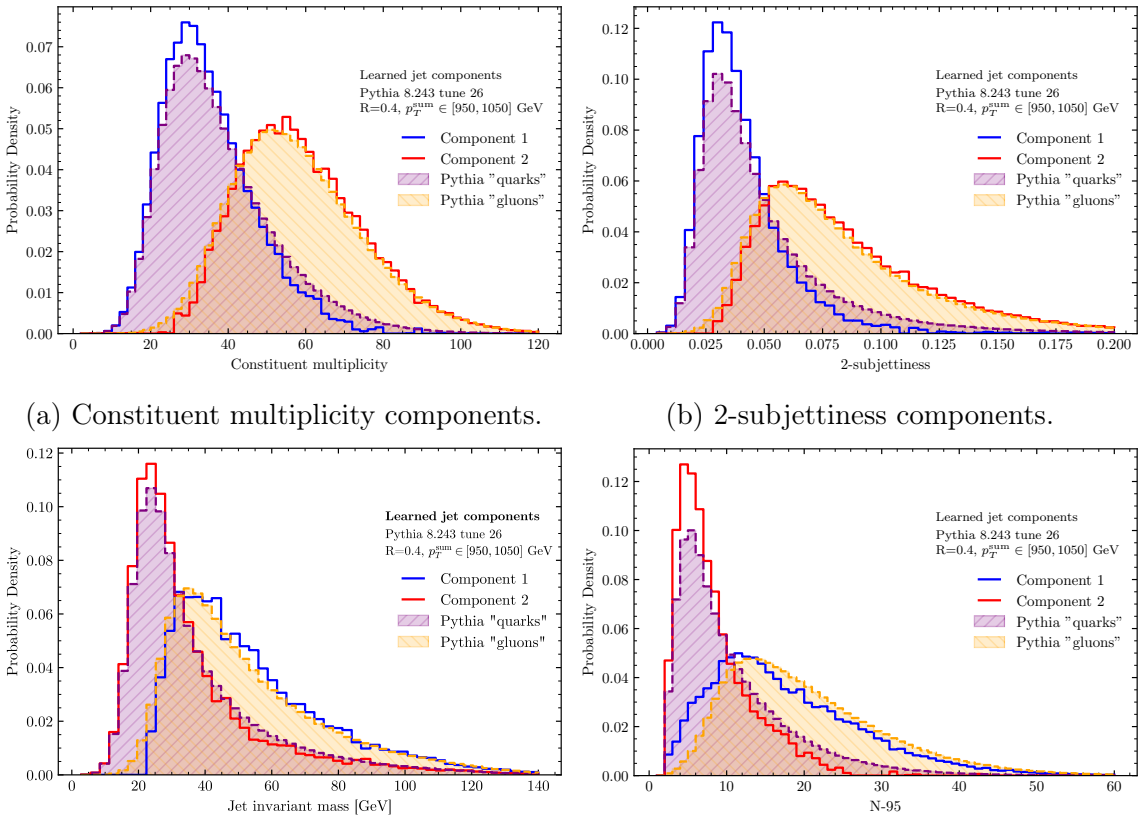
Therefore, the mixing fractions are within approximately 1% of being super-factorized.

7.3.2 Disentangled components

The model is trained using an alternating optimization scheme, using the splitting conic solver [163] implemented in `cvxpy` [164]. All experiments terminate within several hours on a standard laptop computer, suggesting the methods described are quite practical for data analysis in the big-data regime.

In Fig. 7-2, we show the disentangled components learned from applying the topic model to various jet observables from our dijet dataset. Samples are generated by binning across rapidity. Specifically, we apply 5×5 bins across the auxiliary variable pair y_1, y_2. The matrix $\mathbf{D}$ for each is generated by discretizing into 50×50 bins across the observable pair $\mathbf{x}_1, \mathbf{x}_2$. The learned components track the Pythia-generated samples well. This demonstrates that our model has good empirical performance for the quark-gluon tagging problem.

In Fig. 7-3, we show the recovered mixing fractions as a function of the rapidity bin of the hardest jet. The disentangling procedure is performed on constituent



(a) Constituent multiplicity components.

(b) 2-subjettiness components.

(c) Invariant jet mass components.

(d) N-95 components.

Figure 7-2: The components retrieved from factorized topic modeling of dijets. Dijet distributions and ground truth labels were both taken from Pythia simulations. Our method shows good agreement between the learned topics and the ground truth across a variety of jet observables – clockwise from the top left, we show results for constituent multiplicity, 2-subjettiness, invariant jet mass, and N-95.

multiplicity, as this observable has good discriminative power for tagging quarks and gluons. We are able to recover the general trends across the rapidity bins. In particular, forward and backwards jets are quark-enriched, as observed in practice. However, our method overestimates the relative proportion of gluons and underestimates the proportion of quarks, across all rapidity bins. One potential explanation for this behavior is that the true mixing matrix is ill-conditioned. As it is approximately low-rank, finding the optimal mixing matrix given components is highly susceptible to noise compared to finding the components.

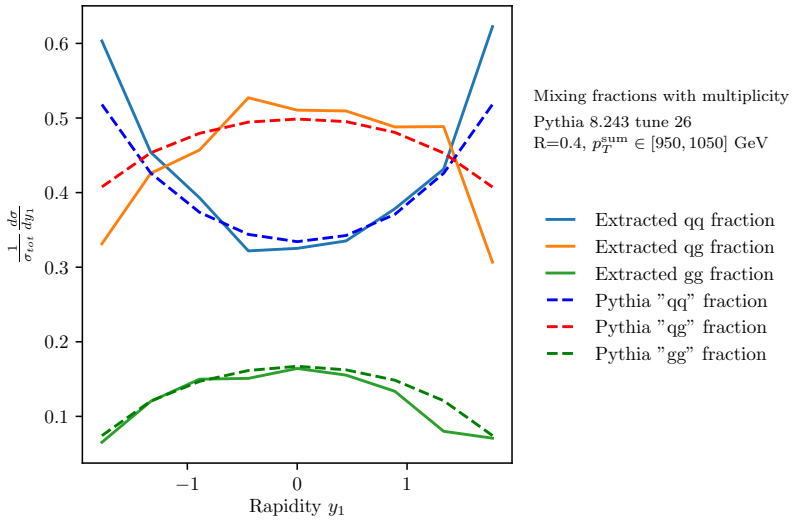


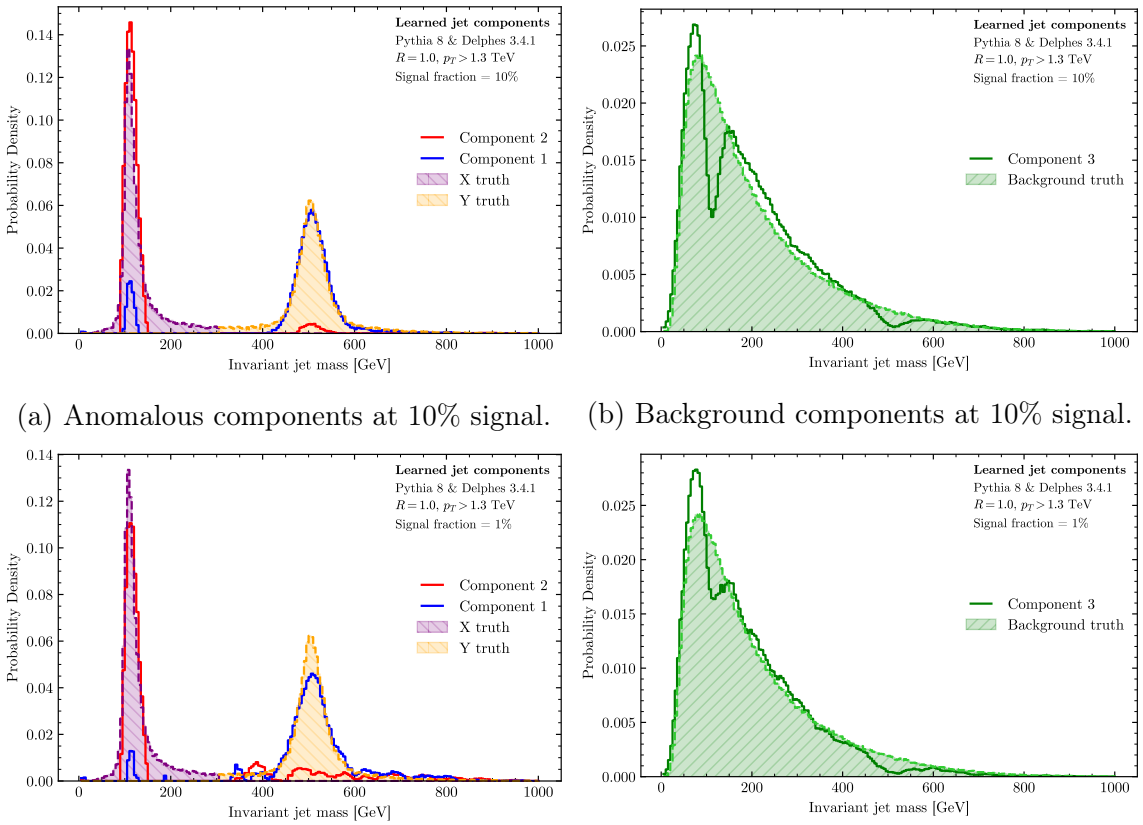
Figure 7-3: The mixing fractions per rapidity bin retrieved from disentangling constituent multiplicity in dijets. We note that the forward and backward bins have a larger proportion of quark jets, as expected from the Pythia labels.

7.4 Reconstructing anomalous resonances

The factorized model we propose is applicable to any sample of dijet events, and therefore is applicable to many other problems beyond quark/gluon discrimination. In this section, we will illustrate it can be used in unsupervised searches for resonant anomalies.

7.4.1 Event generation

We apply our technique to the LHC Olympics 2020 anomaly detection challenge R&D dataset [165]. The events are generated with Pythia 8 [161] and Delphes 3.4.1 [166], with no pileup or multi-parton interactions. The signal process in this dataset is a hypothetical W' boson with mass $m_{W'} = 3.5$ TeV, which decays into X, Y bosons with masses $m_X = 500$ GeV and $m_Y = 100$ GeV, respectively. The boosted bosons decay into quark pairs, which are clustered together with the FastJet [162] implementation of the anti-k_T algorithm with $R = 1.0$. The wide radius of the clustering algorithm captures all the decay products of the quark pair within a single jet, meaning that the resulting event looks like a dijet. A trigger of $p_T^{\text{leading}} > 1.2$ TeV is applied, yielding



(a) Anomalous components at 10% signal.

(b) Background components at 10% signal.

(c) Anomalous components at 1% signal.

(d) Background components at 1% signal.

Figure 7-4: The components retrieved from factorized topic modeling of the LHC Olympics R&D dataset. Our method shows good agreement between the learned topics and the ground truth on the invariant jet mass observable. We are able to recover both of the resonant masses (at 100 GeV and 500 GeV) with signal fraction of 10% (top row) and 1% (bottom row), up to mutual irreducibility.

1,000,000 background QCD dijets, with an additional 100,000 signal events.

7.4.2 Results and sensitivity

In Fig. 7-4, we demonstrate the performance of our algorithm in recovering the mass distributions for the dijets in the anomaly detection dataset. We learn a model with 3 topics, corresponding to $\mathfrak{p}_X, \mathfrak{p}_Y, \mathfrak{p}_{\text{QCD}}$, respectively. To generate these figures, we consider a signal fraction of 10%, and 1% respectively, solve the topic model, and then re-weight the component distributions by subtracting the overall background distribution and renormalizing. This is necessary because the distributions are only accurate

up to mutual reducibility, as the true components do not satisfy the separability hypothesis. An artifact of this reducibility procedure is visible in the background distribution, which exhibits noticeable dips around the resonant masses. In addition, one or both of the components show small peaks at both resonant masses. Again, these artifacts are expected due to the lack of separability. As the resonances appear on the histogram as bumps in a smoothly falling background, there are no regions within the resonance mass that are only supported in the anomalous component distributions. Our model does not contain a prior on the number of modes of the component, or any smoothness assumptions. In addition, the algorithms used to optimize the model return extremal points in the polytope of all feasible solutions, as they are alternating convex projections that remain on the boundary of the constraint set. Therefore, the solution will force the components to be as orthogonal as possible.

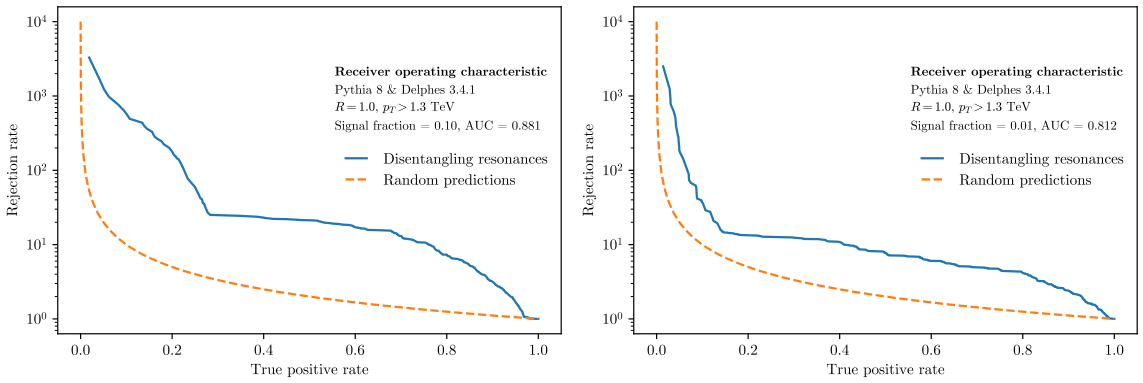
After this processing step, in both cases, we are able to recover the correct masses of 100 and 500 GeV respectively. We note that the noise in the recovered distributions is noticeably larger at the lower signal fractions, as expected. However, in both cases, our model has significant discriminative power. In particular, the model can infer which process any event was generated from using the likelihood ratio:

$$\mathcal{L}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathsf{f}(\text{signal}) \cdot \mathsf{p}_{\text{signal}}(\mathbf{x}_1, \mathbf{x}_2)}{\mathsf{f}(\text{bg}) \cdot \mathsf{p}_{\text{bg}}(\mathbf{x}_1, \mathbf{x}_2)} \tag{7.23}$$

$$= \frac{\mathsf{f}(\mathsf{X}, \mathsf{Y})\mathsf{p}_{\mathsf{X}}(\mathbf{x}_1)\mathsf{p}_{\mathsf{Y}}(\mathbf{x}_2) + \mathsf{f}(\mathsf{Y}, \mathsf{X})\mathsf{p}_{\mathsf{Y}}(\mathbf{x}_1)\mathsf{p}_{\mathsf{X}}(\mathbf{x}_2)}{\mathsf{f}(\text{QCD, QCD}) \cdot \mathsf{p}_{\text{QCD}}(\mathbf{x}_1)\mathsf{p}_{\text{QCD}}(\mathbf{x}_2)}. \tag{7.24}$$

Using this likelihood ratio as a discriminant, we can test the ability of our model to classify events relative to the ground truth in the dataset. In Fig. 7-5, we show the receiver operating characteristic curve for the 10% and 1% signal fraction. In both cases, the model performs very well compared to randomness, with AUCs of 0.88 and 0.81, respectively.

To test the sensitivity of the model, we report the AUC of the model while varying the signal fraction from 10% down to 0.1% in 7-6. The model is initialized 30 times with different parameters, and we show the value of the AUC to one standard error. The performance of the model is strong even at low signal fractions.



(a) ROC curve, signal fraction $\mathtt{f_{sg}} = 10\%$. (b) ROC curve, signal fraction $\mathtt{f_{sg}} = 1\%$.

Figure 7-5: The receiver operating characteristic curve recovered from disentangling resonant masses. At both 10% and 1% signal fraction, almost all the anomalies are identified with a component-based likelihood ratio test.

7.5 Next steps

Topic models have shown promise for understanding the generative processes behind jets at the LHC [136, 138]. In this work, our goal was to leverage the factorization theorem for jet substructure to design more powerful models. To that end, we have introduced a new factorized topic model to disentangle dijets. The model is unsupervised, meaning that it does not rely on ground-truth labels for any events within the sample. It is also non-parametric, assuming no form *a priori* for the distribution over jet observables. Our model requires fairly few assumptions – we only require the mixture to be full-rank, and for the true components to be non-degenerate to retrieve and we have demonstrated its applicability to both quark/gluon discrimination, and anomaly detection in a BSM resonance search. In summary, we have shown that modeling dijets as a factorized mixture of components is a promising direction for analysis in multiple relevant problems in jet physics. Our model has significant discriminative power for low-signal anomalies, can perform background density estimation, and is able to discriminate between quark and gluon jets, in a fully data-driven and unsupervised manner.

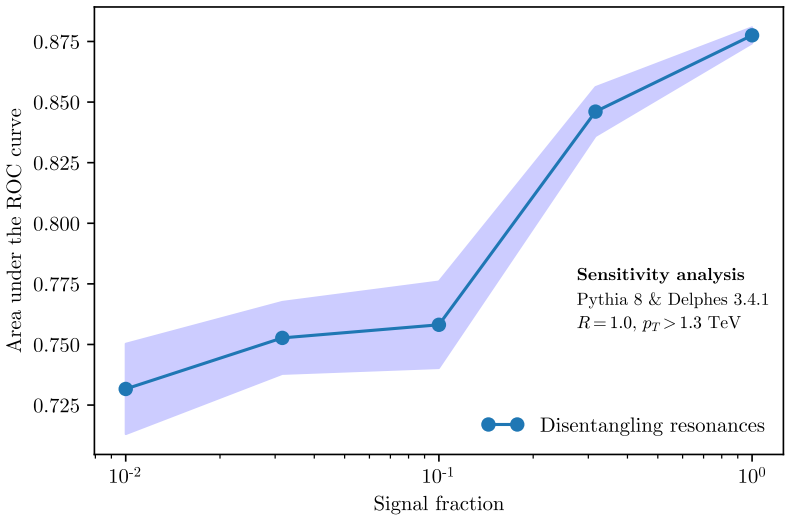


Figure 7-6: The AUC recovered from disentangling resonant masses at different signal fractions. The model has discriminative power down to the regime of 0.1% signal, corresponding to 1,000 signal events interspersed through 1,000,000 background events.

Chapter 8

Conclusions

In this work, we have made steps towards a better understanding of anomaly detection at particle colliders. The data collected by colliders consists of event signatures, which represent energy distributions over the calorimeters around the inner ring of the accelerator. These signatures can be though of as empirical approximations to some continuous, underlying measure. By operating on these signatures as if they are distributions, we can induce a metric on the space of all collider events. This allows us to better understand the similarities and differences between events. We base our metric on the theory of optimal transport, as transport distances have certain beneficial theoretical guarantees related to empirical distributions. Further, they satisfy key physical properties like infrared and collinear safety, making them robust to certain degeneracies in quantum field theory.

Next, we used kernels based on the sliced Wasserstein distance, a variant of optimal transport that induces a positive definite Gaussian kernel, for anomaly detection. From a theoretical perspective, we have demonstrated an unbiased sliced Wasserstein kernel, and proven that it enjoys certain geometric benefits over its exact Wasserstein counterpart. In addition, we have shown that the centroidal Voronoi tessellation is a good approximation with respect to quadrature error for a broad class of functions, including the sliced Wasserstein kernel. Experimentally, we have built a framework for understanding which types of models are well-suited for certain classes of anomalies. We empirically demonstrate the performance of these discriminative models using

our transport-based kernels on two real-world anomaly detection tasks. Finally, we have built a generative model that encodes the factorization theorem as a statistical constraint, and performs extremely well on real-world tasks like quark/gluon discrimination and resonant anomaly detection.

So, what's next? Our research leaves many questions open and suggests several directions for further research.

1. **Kernels and approximations.** Our framework for analyzing subsampling error through quadrature is a new approach to the problem, and suggests several potential extensions.

    - *Rigorous bounds on SVMs.* We leave open the quadrature analysis for specific error functions. A future work might seek to complete this analysis for the SVM objective function, and seek to quantify the difference in the actual mean squared error. This would be, to our knowledge, a new result for the SVM.

    - *Barycenters.* Solidifying the link between the Wasserstein barycenter problem and Voronoi tessellations could suggest more efficient algorithms for computing barycenters explicitly. In addition, bounding the Wasserstein error due to farthest point sampling using our framework would be an interesting result, as fast algorithms for farthest point sampling already exist.

2. **Wasserstein-type kernels.** Our unbiased and fast kernels have good theoretical guarantees but perform poorly in practice, and understanding this gap is of interest.

    - *Monte Carlo Wasserstein.* Understanding why the variance of this estimator is so high, and finding a low-variance approximation would allow for fast computation of the kernel in practice. Further, generalizing this type of estimator to the exact and entropic Wasserstein cases would yield immense practical benefit.

3. **Discriminative anomaly detection.** We showed that the discriminative methods perform well on certain types of anomalies, but not others; additionally, we suggest why this may be the case. However, a method that works across anomaly types is still unknown.

    - *Better-designed models.* Finding a loss function that can optimize for multiple types of anomalies with a single minimization procedure would be of practical interest.
    - *Wasserstein kernels as features.* One potential anomaly detection method that might work is to use the Wasserstein kernel between a datapoint and a cluster centroid of the dataset as a feature vector, and try and isolate anomalies that are "far away" from the

4. **Generative models.** Within the framework of generative anomaly detection, we have restricted our focus to a specific set of observables and assumptions. Future research could seek to expand beyond these restrictions, along the following directions.

    - *Tensor factorization.* We have restricted ourselves to dijets in a single observable; however, theoretically our method is generalizable to any number of jets. Finding efficient algorithms to solve the analogous minimization problem in practice is an open question.
    - *Separability.* We have assumed that separability is necessary to recover mutually irreducible components, but recent work has shown that identifiability does not require separability for certain classes of topic models. Extending these models to the factorized picture would be quite relevant, as it would allow us to understand the mixing fractions without reweighting.
    - *Neural topic models.* It may be possible to simultaneously train a topic model and learn the observable on which it is trained by leveraging backpropagation. Understanding the behavior of topic models on learned ob-

servables, like Energy Flow Polynomials [167] would allow for a deeper understanding of jet substructure.

Machine learning is a powerful tool, when wielded properly, and it only becomes more powerful when it is imbued with the proper statistical and physical structure corresponding to the task at hand. The results we have achieved through optimal transport, generative modelling, and kernel methods demonstrate the promise of geometric learning techniques when applied to event signatures at colliders.

Bibliography

[1] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. The anti-ktjet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063–063, Apr 2008. ISSN 1029-8479. doi: 10.1088/1126-6708/2008/04/063. URL http://dx.doi.org/10.1088/1126-6708/2008/04/063.

[2] C. Grojean. Higgs physics, 2017.

[3] Abdelhak Djouadi. The anatomy of electroweak symmetry breaking. *Physics Reports*, 457(1-4):1–216, Feb 2008. ISSN 0370-1573. doi: 10.1016/j.physrep.2007.10.004. URL http://dx.doi.org/10.1016/j.physrep.2007.10.004.

[4] G. Aad, T. Abajyan, B. Abbott, J. Abdallah, S. Abdel Khalek, A.A. Abdelalim, O. Abdinov, R. Aben, B. Abi, M. Abolins, and et al. Observation of a new particle in the search for the standard model higgs boson with the atlas detector at the lhc. *Physics Letters B*, 716(1):1–29, Sep 2012. ISSN 0370-2693. doi: 10.1016/j.physletb.2012.08.020. URL http://dx.doi.org/10.1016/j.physletb.2012.08.020.

[5] Aude Genevay. *Entropy-regularized optimal transport for machine learning*. PhD thesis, 2019.

[6] Christian W Bauer, Dan Pirjol, and Iain W Stewart. Soft-collinear factorization in effective field theory. *Physical Review D*, 65(5):054022, 2002.

[7] Silvan S. Schweber and Felix M. H. Villars. Qed and the men who made it: Dyson, feynman, schwinger, and tomonaga. *American Journal of Physics*, 63 (4):383–384, 1995. doi: 10.1119/1.17926. URL https://doi.org/10.1119/1.17926.

[8] Benjamin Nachman. Investigating the quantum properties of jets and the search for a supersymmetric top quark partner with the atlas detector, 2016.

[9] Sundance O Bilson-Thompson, Fotini Markopoulou, and Lee Smolin. Quantum gravity and the standard model. *Classical and Quantum Gravity*, 24(16): 3975–3993, Jul 2007. ISSN 1361-6382. doi: 10.1088/0264-9381/24/16/002. URL http://dx.doi.org/10.1088/0264-9381/24/16/002.

[10] A. Zee. *Quantum field theory in a nutshell*. 11 2003. ISBN 978-0-691-14034-6.

[11] M. Tanabashi et al. Review of particle physics. *Phys. Rev. D*, 98:030001, Aug 2018. doi: 10.1103/PhysRevD.98.030001. URL https://link.aps.org/doi/10.1103/PhysRevD.98.030001.

[12] David Griffiths. *Introduction to elementary particles*. 2008. ISBN 978-3-527-40601-2.

[13] David J. Gross and Frank Wilczek. Ultraviolet behavior of non-abelian gauge theories. *Phys. Rev. Lett.*, 30:1343–1346, Jun 1973. doi: 10.1103/PhysRevLett.30.1343. URL https://link.aps.org/doi/10.1103/PhysRevLett.30.1343.

[14] A. Di Giacomo. Understanding color confinement. *EPJ Web of Conferences*, 70:00019, 2014. ISSN 2100-014X. doi: 10.1051/epjconf/20147000019. URL http://dx.doi.org/10.1051/epjconf/20147000019.

[15] Kenneth G. Wilson. Confinement of quarks. *Phys. Rev. D*, 10:2445–2459, Oct 1974. doi: 10.1103/PhysRevD.10.2445. URL https://link.aps.org/doi/10.1103/PhysRevD.10.2445.

[16] B. Andersson, S. Mohanty, and F. Söderberg. The lund fragmentation process for a multi-gluon string according to the area law. *The European Physical Journal C*, 21(4):631–647, Jul 2001. ISSN 1434-6052. doi: 10.1007/s100520100757. URL http://dx.doi.org/10.1007/s100520100757.

[17] Victor Coco, Pierre-Antoine Delsart, Juan Rojo-Chacon, Gregory Soyez, and Christian Sander. Jets and jet algorithms. In *Proceedings, HERA and the LHC Workshop Series on the implications of HERA for LHC physics: 2006-2008*, pages 182–204, 3 2009. doi: 10.3204/DESY-PROC-2009-02/54.

[18] Andrew J. Larkoski, Ian Moult, and Benjamin Nachman. Jet substructure at the large hadron collider: A review of recent advances in theory and machine learning. *Physics Reports*, 841:1 – 63, 2020. ISSN 0370-1573. doi: https://doi.org/10.1016/j.physrep.2019.11.001. URL http://www.sciencedirect.com/science/article/pii/S0370157319303643. Jet substructure at the Large Hadron Collider: A review of recent advances in theory and machine learning.

[19] Jesse Thaler and Ken Van Tilburg. Identifying boosted objects with n-subjettiness. *Journal of High Energy Physics*, 2011(3), Mar 2011. ISSN 1029-8479. doi: 10.1007/jhep03(2011)015. URL http://dx.doi.org/10.1007/JHEP03(2011)015.

[20] V. Khachatryan, A.M. Sirunyan, A. Tumasyan, W. Adam, T. Bergauer, M. Dragicevic, J. Erö, C. Fabjan, M. Friedl, R. Frühwirth, and et al. Measurements of jet multiplicity and differential production cross sections of z+jets events in proton-proton collisions at s=7 tev. *Physical Review D*, 91 (5), Mar 2015. ISSN 1550-2368. doi: 10.1103/physrevd.91.052008. URL http://dx.doi.org/10.1103/PhysRevD.91.052008.

[21] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. Energy flow polynomials: a complete linear basis for jet substructure. *Journal of High Energy Physics*, 2018(4), Apr 2018. ISSN 1029-8479. doi: 10.1007/jhep04(2018)013. URL http://dx.doi.org/10.1007/JHEP04(2018)013.

[22] Daniele Bertolini, Tucker Chan, and Jesse Thaler. Jet observables without jet algorithms. *Journal of High Energy Physics*, 2014(4), Apr 2014. ISSN 1029-8479. doi: 10.1007/jhep04(2014)013. URL http://dx.doi.org/10.1007/JHEP04(2014)013.

[23] Andrew J. Larkoski and Eric M. Metodiev. A Theory of Quark vs. Gluon Discrimination. *JHEP*, 10:014, 2019. doi: 10.1007/JHEP10(2019)014.

[24] Andrew J. Larkoski, Gavin P. Salam, and Jesse Thaler. Energy Correlation Functions for Jet Substructure. *JHEP*, 06:108, 2013. doi: 10.1007/JHEP06(2013)108.

[25] Andrew J. Larkoski, Ian Moult, and Benjamin Nachman. Jet Substructure at the Large Hadron Collider: A Review of Recent Advances in Theory and Machine Learning. *Phys. Rept.*, 841:1–63, 2020. doi: 10.1016/j.physrep.2019.11.001.

[26] Andrew J. Larkoski, Ian Moult, and Duff Neill. Power Counting to Better Jet Observables. *JHEP*, 12:009, 2014. doi: 10.1007/JHEP12(2014)009.

[27] Mrinal Dasgupta, Frédéric Dreyer, Gavin P. Salam, and Gregory Soyez. Small-radius jets to all orders in QCD. *JHEP*, 04:039, 2015. doi: 10.1007/JHEP04(2015)039.

[28] D. Adams et al. Towards an Understanding of the Correlations in Jet Substructure. *Eur. Phys. J. C*, 75(9):409, 2015. doi: 10.1140/epjc/s10052-015-3587-2.

[29] Ian Moult, Lina Necib, and Jesse Thaler. New Angles on Energy Correlation Functions. *JHEP*, 12:153, 2016. doi: 10.1007/JHEP12(2016)153.

[30] Andrew J. Larkoski, Jesse Thaler, and Wouter J. Waalewijn. Gaining (Mutual) Information about Quark/Gluon Discrimination. *JHEP*, 11:129, 2014. doi: 10.1007/JHEP11(2014)129.

[31] Katherine Fraser and Matthew D. Schwartz. Jet Charge and Machine Learning. *JHEP*, 10:093, 2018. doi: 10.1007/JHEP10(2018)093.

[32] Mrinal Dasgupta, Kamel Khelifa-Kerfa, Simone Marzani, and Michael Spannowsky. On jet mass distributions in Z+jet and dijet processes at the LHC. *JHEP*, 10:126, 2012. doi: 10.1007/JHEP10(2012)126.

[33] *Les Houches 2017: Physics at TeV Colliders Standard Model Working Group Report*, 3 2018.

[34] Roman Kogler et al. Jet Substructure at the Large Hadron Collider: Experimental Review. *Rev. Mod. Phys.*, 91(4):045003, 2019. doi: 10.1103/RevModPhys.91.045003.

[35] Albert M Sirunyan et al. Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements. *Eur. Phys. J. C*, 80 (1):4, 2020. doi: 10.1140/epjc/s10052-019-7499-4.

[36] A. Altheimer et al. Jet Substructure at the Tevatron and LHC: New results, new tools, new benchmarks. *J. Phys. G*, 39:063001, 2012. doi: 10.1088/0954-3899/39/6/063001.

[37] Georges Aad et al. Performance of jet substructure techniques for large-R jets in proton-proton collisions at $\sqrt{\mathsf{s}} = 7$ TeV using the ATLAS detector. *JHEP*, 09:076, 2013. doi: 10.1007/JHEP09(2013)076.

[38] Georges Aad et al. Jet mass and substructure of inclusive jets in $\sqrt{\mathsf{s}} = 7$ TeV pp collisions with the ATLAS experiment. *JHEP*, 05:128, 2012. doi: 10.1007/JHEP05(2012)128.

[39] Performance of large-R jets and jet substructure reconstruction with the ATLAS detector. 7 2012.

[40] Olmo Cerri, Thong Q. Nguyen, Maurizio Pierini, Maria Spiropulu, and Jean-Roch Vlimant. Variational Autoencoders for New Physics Mining at the Large Hadron Collider. *arXiv e-prints*, art. arXiv:1811.10276, Nov 2018.

[41] Marco Farina, Yuichiro Nakai, and David Shih. Searching for New Physics with Deep Autoencoders. *arXiv e-prints*, art. arXiv:1808.08992, Aug 2018.

[42] Jack H Collins, Kiel Howe, and Benjamin Nachman. Extending the Bump Hunt with Machine Learning. *arXiv e-prints*, art. arXiv:1902.02634, Feb 2019.

[43] Oz Amram and Cristina Mantilla Suarez. Tag n' train: A technique to train improved classifiers on unlabeled data, 2020.

[44] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.

[45] Cédric Villani. *Topics in Optimal Transportation*. Number 58. American Mathematical Soc., 2003.

[46] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. The Metric Space of Collider Events. *arXiv e-prints*, art. arXiv:1902.02346, Feb 2019.

[47] I Csabai, F Czakó, and Z Fodor. Combined neural network—qcd classifier for quark and gluon jet separation. *Nuclear Physics B*, 374(2):288–308, 1992.

[48] Leif Lönnblad, Carsten Peterson, and Thorsteinn Rögnvaldsson. Using neural networks to identify jets. *Nuclear Physics B*, 349(3):675–702, 1991.

[49] Patrick T Komiske, Eric M Metodiev, and Matthew D Schwartz. Deep learning in color: towards automated quark/gluon jet discrimination. *Journal of High Energy Physics*, 2017(1):110, 2017.

[50] Taoli Cheng. Recursive neural networks in quark/gluon tagging. *Computing and Software for Big Science*, 2(1):3, 2018.

[51] Jason Gallicchio and Matthew D. Schwartz. Pure samples of quark and gluon jets at the lhc. *Journal of High Energy Physics*, 2011(10), Oct 2011. ISSN 1029-8479. doi: 10.1007/jhep10(2011)103. URL http://dx.doi.org/10.1007/JHEP10(2011)103.

[52] Andrew J Larkoski, Gavin P Salam, and Jesse Thaler. Energy correlation functions for jet substructure. *Journal of High Energy Physics*, 2013(6):108, 2013.

[53] Eric M. Metodiev and Jesse Thaler. Jet topics: Disentangling quarks and gluons at colliders. *Physical Review Letters*, 120(24), Jun 2018. ISSN 1079-7114. doi: 10.1103/physrevlett.120.241602. URL http://dx.doi.org/10.1103/PhysRevLett.120.241602.

[54] Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pages 146–158, 1975.

[55] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[56] Alfred Galichon. *Optimal transport methods in economics*. Princeton University Press, 2018.

[57] Filippo Santambrogio. Models and applications of optimal transport in economics, traffic and urban planning. *arXiv preprint arXiv:1009.3857*, 2010.

[58] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2): 99–121, 2000.

[59] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):1–11, 2015.

[60] Julien Rabin and Nicolas Papadakis. Convex color image segmentation with optimal transport distances. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 256–269. Springer, 2015.

[61] Steven Haker, Lei Zhu, Allen Tannenbaum, and Sigurd Angenent. Optimal mass transport for registration and warping. *International Journal of computer vision*, 60(3):225–240, 2004.

[62] F. M. Ngolè Mboula and J. L. Starck. Psf field learning based on optimal transport distances, 2017.

[63] Morgan A. Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngolè, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, Jan 2018. ISSN 1936-4954. doi: 10.1137/17m1140431. URL http://dx.doi.org/10.1137/17M1140431.

[64] Max Sommerfeld, Jörn Schrieber, Yoav Zemel, and Axel Munk. Optimal transport: Fast probabilistic approximation with exact solvers, 2018.

[65] Eustasio Del Barrio and Jean-Michel Loubes. Central limit theorem for empirical transportation cost in general dimension, 2017.

[66] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013.

[67] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[68] Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*, pages 4543–4553, 2019.

[69] Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration, 2017.

[70] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. Metric space of collider events. *Physical Review Letters*, 123(4), Jul 2019. ISSN 1079-7114. doi: 10.1103/physrevlett.123.041801. URL http://dx.doi.org/10.1103/PhysRevLett.123.041801.

[71] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. The hidden geometry of particle collisions, 2020.

[72] Soheil Kolouri, Yang Zou, and Gustavo K Rohde. Sliced Wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5258–5267, 2016.

[73] Bernhard Schölkopf. The kernel trick for distances. In *Advances in neural information processing systems*, pages 301–307, 2001.

[74] Manqi Zhao and Venkatesh Saligrama. Anomaly detection with score functions based on nearest neighbor graphs. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2250–2258. Curran Associates, Inc., 2009.

[75] Krikamol Muandet and Bernhard Schölkopf. One-Class Support Measure Machines for Group Anomaly Detection. *arXiv e-prints*, art. arXiv:1303.0309, Mar 2013.

[76] Christopher JC Burges, Bernhard Scholkopf, and Alexander J Smola. *Advances in kernel methods: support vector learning*. MIT press Cambridge, MA, USA:, 1999.

[77] Aasa Feragen, Francois Lauze, and Soren Hauberg. Geodesic exponential kernels: When curvature and linearity conflict. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3032–3042, 2015.

[78] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2008.

[79] Charless Fowlkes, Serge Belongie, Fan Chung, and Jitendra Malik. Spectral grouping using the nystrom method. *IEEE transactions on pattern analysis and machine intelligence*, 26(2):214–225, 2004.

[80] Christopher KI Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *Advances in neural information processing systems*, pages 682–688, 2001.

[81] Dino Oglic and Thomas Gärtner. Nyström method with kernel k-means++ samples as landmarks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2652–2660. JMLR. org, 2017.

[82] Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Fast dpp sampling for nystr\" om with application to kernel methods. *arXiv preprint arXiv:1603.06052*, 2016.

[83] Kai Zhang and James T Kwok. Clustered nyström method for large scale manifold learning and dimension reduction. *IEEE Transactions on Neural Networks*, 21(10):1576–1587, 2010.

[84] Francis Bach. Sharp analysis of low-rank kernel matrix approximations. In *Conference on Learning Theory*, pages 185–209, 2013.

[85] Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017.

[86] Ayoub Belhadji, Rémi Bardenet, and Pierre Chainais. Kernel quadrature with dpps, 2019.

[87] Yang Liu, Wenping Wang, Bruno Lévy, Feng Sun, Dong-Ming Yan, Lin Lu, and Chenglei Yang. On centroidal voronoi tessellation—energy smoothness and fast computation. *ACM Trans. Graph.*, 28(4), September 2009. ISSN 0730-0301. doi: 10.1145/1559755.1559758. URL https://doi.org/10.1145/1559755.1559758.

[88] Qiang Du and Tak-Win Wong. Numerical studies of macqueen's k-means algorithm for computing the centroidal voronoi tessellations. *Computers & Mathematics with Applications*, 44(3-4):511–523, 2002.

[89] Franz Aurenhammer. Power diagrams: properties, algorithms and applications. *SIAM Journal on Computing*, 16(1):78–96, 1987.

[90] Sebastian Claici, Edward Chien, and Justin Solomon. Stochastic wasserstein barycenters. *arXiv preprint arXiv:1802.05757*, 2018.

[91] Yuki Saka, Max Gunzburger, and John Burkardt. Latinized, improved lhs, and cvt point sets in hypercubes.

[92] Qiang Du and Desheng Wang. The optimal centroidal voronoi tessellations and the gersho's conjecture in the three-dimensional space. *Computers & Mathematics with Applications*, 49(9-10):1355–1373, 2005.

[93] Krikamol Muandet and Bernhard Schölkopf. One-Class Support Measure Machines for Group Anomaly Detection. *arXiv e-prints*, art. arXiv:1303.0309, Mar 2013.

[94] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *ICLR 2018*, 2018.

[95] Xiao Yang, Dafang He, Zihan Zhou, Daniel Kifer, and C Lee Giles. Learning to read irregular text with attention mechanisms. In *IJCAI*, volume 1, page 3, 2017.

[96] Rainer Kelz and Gerhard Widmer. Towards interpretable polyphonic transcription with invertible neural networks. *International Society for Music Information Retrieval Conference*, 2019.

[97] Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced Wasserstein discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10285–10295, 2019.

[98] Zhixin Shu, Sunil Hadap, Eli Shechtman, Kalyan Sunkavalli, Sylvain Paris, and Dimitris Samaras. Portrait lighting transfer using a mass transport approach. *ACM Transactions on Graphics (TOG)*, 36(4):1, 2017.

[99] Mathieu Carrière, Marco Cuturi, and Steve Oudot. Sliced Wasserstein kernel for persistence diagrams. In *International Conference on Machine Learning*, pages 664–673. JMLR. org, 2017.

[100] Peter W Glynn and Chang-Han Rhee. Exact estimation for Markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A):377–389, 2014.

[101] Anne-Marie Lyne, Mark Girolami, Yves Atchadé, Heiko Strathmann, Daniel Simpson, et al. On Russian roulette estimates for bayesian inference with doubly-intractable likelihoods. *Statistical Science*, 30(4):443–467, 2015.

[102] Philippe Berthet, Jean-Claude Fort, and Thierry Klein. A central limit theorem for Wasserstein type distances between two different laws. *arXiv preprint arXiv:1710.09763*, 2017.

[103] Fred J Hickernell, Christiane Lemieux, Art B Owen, et al. Control variates for quasi-monte carlo. *Statistical Science*, 20(1):1–31, 2005.

[104] Michael B Giles. Multilevel monte carlo methods. In *Monte Carlo and Quasi-Monte Carlo Methods 2012*, pages 83–103. Springer, 2013.

[105] Arun Jambulapati, Aaron Sidford, and Kevin Tian. A direct o(1/e) iteration parallel algorithm for optimal transport. *ArXiv Preprint*, 2019:2, 1906.

[106] Tianyi Lin, Nhat Ho, and Michael I. Jordan. On the efficiency of the sinkhorn and greenkhorn algorithms and their acceleration for optimal transport, 2019.

[107] Richard Mansfield Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

[108] Eric M. Metodiev and Jesse Thaler. Jet Topics: Disentangling Quarks and Gluons at Colliders. *Phys. Rev. Lett.*, 120(24):241602, 2018. doi: 10.1103/PhysRevLett.120.241602.

[109] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011.

[110] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141, 2017.

[111] R Flamary. Pot: Python optimal transport. 2017.

[112] Nima Aghaeepour, Greg Finak, Holger Hoos, Tim R Mosmann, Ryan Brinkman, Raphael Gottardo, Richard H Scheuermann, FlowCAP Consortium, Dream Consortium, et al. Critical assessment of automated flow cytometry data analysis techniques. *Nature methods*, 10(3):228, 2013.

[113] Gregor Kasieczka, Tilman Plehn, Michael Russell, and Torben Schell. Deep-learning Top Taggers or The End of QCD? *JHEP*, 05:006, 2017. doi: 10.1007/JHEP05(2017)006.

[114] Christopher T Hill. Topcolor assisted technicolor. *arXiv preprint hep-ph/9411426*, 1994.

[115] Alfred O Hero. Geometric entropy minimization (gem) for anomaly detection and localization. In *Advances in Neural Information Processing Systems*, pages 585–592, 2007.

[116] Kumar Sricharan and Alfred O Hero. Efficient anomaly detection using bipartite k-nn graphs. In *Advances in Neural Information Processing Systems*, pages 478–486, 2011.

[117] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[118] Jason Gallicchio and Matthew D. Schwartz. Pure Samples of Quark and Gluon Jets at the LHC. *JHEP*, 10:103, 2011. doi: 10.1007/JHEP10(2011)103.

[119] Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. An operational definition of quark and gluon jets. *JHEP*, 11:059, 2018. doi: 10.1007/JHEP11(2018)059.

[120] Eric M. Metodiev, Benjamin Nachman, and Jesse Thaler. Classification without labels: Learning from mixed samples in high energy physics. *JHEP*, 10:174, 2017. doi: 10.1007/JHEP10(2017)174.

[121] Patrick T. Komiske, Eric M. Metodiev, Benjamin Nachman, and Matthew D. Schwartz. Learning to classify from impure samples with high-dimensional data. *Phys. Rev. D*, 98(1):011502, 2018. doi: 10.1103/PhysRevD.98.011502.

[122] Anders Andreassen, Benjamin Nachman, and David Shih. Simulation Assisted Likelihood-free Anomaly Detection. 1 2020.

[123] Benjamin Nachman and David Shih. Anomaly Detection with Density Estimation. 1 2020.

[124] Jack H. Collins, Kiel Howe, and Benjamin Nachman. Anomaly Detection for Resonant New Physics with Machine Learning. *Phys. Rev. Lett.*, 121(24):241803, 2018. doi: 10.1103/PhysRevLett.121.241803.

[125] The ATLAS collaboration. Search for resonances decaying to photon pairs in 3.2 fb^{-1} of pp collisions at $\sqrt{\mathrm{s}} = 13$ TeV with the ATLAS detector. 12 2015.

[126] Georges Aad et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B*, 716:1–29, 2012. doi: 10.1016/j.physletb.2012.08.020.

[127] Anja Butter, Gregor Kasieczka, Tilman Plehn, and Michael Russell. Deep-learned Top Tagging with a Lorentz Layer. *SciPost Phys.*, 5(3):028, 2018. doi: 10.21468/SciPostPhys.5.3.028.

[128] Leandro G. Almeida, Mihailo Backović, Mathieu Cliche, Seung J. Lee, and Maxim Perelstein. Playing Tag with ANN: Boosted Top Identification with Pattern Recognition. *JHEP*, 07:086, 2015. doi: 10.1007/JHEP07(2015)086.

[129] Chase Shimmin, Peter Sadowski, Pierre Baldi, Edison Weik, Daniel Whiteson, Edward Goul, and Andreas Søgaard. Decorrelated Jet Substructure Tagging using Adversarial Neural Networks. *Phys. Rev. D*, 96(7):074034, 2017. doi: 10.1103/PhysRevD.96.074034.

[130] T. Aaltonen, A. Buzatu, B. Kilminster, Y. Nagai, and W. Yao. Improved b-jet Energy Correction for $\mathrm{H} \to \mathrm{b\bar{b}}$ Searches at CDF. 7 2011.

[131] Morad Aaboud et al. Performance of top-quark and W-boson tagging with ATLAS in Run 2 of the LHC. *Eur. Phys. J. C*, 79(5):375, 2019. doi: 10.1140/epjc/s10052-019-6847-8.

[132] The ATLAS collaboration. Performance of Top Quark and W Boson Tagging in Run 2 with ATLAS. 8 2017.

[133] Serguei Chatrchyan et al. Search for Resonances in the Dijet Mass Spectrum from 7 TeV pp Collisions at CMS. *Phys. Lett. B*, 704:123–142, 2011. doi: 10.1016/j.physletb.2011.09.015.

[134] Georges Aad et al. Search for New Physics in the Dijet Mass Distribution using 1 fb^{-1} of pp Collision Data at $\sqrt{\mathrm{s}}$ =7 TeV collected by the ATLAS Detector. *Phys. Lett. B*, 708:37–54, 2012. doi: 10.1016/j.physletb.2012.01.035.

[135] Oriol Domenech, Alex Pomarol, and Javi Serra. Probing the SM with Dijets at the LHC. *Phys. Rev. D*, 85:074030, 2012. doi: 10.1103/PhysRevD.85.074030.

[136] Eric M. Metodiev and Jesse Thaler. Jet Topics: Disentangling Quarks and Gluons at Colliders. *Phys. Rev. Lett.*, 120(24):241602, 2018. doi: 10.1103/PhysRevLett.120.241602.

[137] R.Keith Ellis, W.James Stirling, and B.R. Webber. *QCD and collider physics*, volume 8. Cambridge University Press, 2 2011. ISBN 978-0-511-82328-2, 978-0-521-54589-1.

[138] Georges Aad et al. Properties of jet fragmentation using charged particles measured with the ATLAS detector in pp collisions at $\sqrt{\mathrm{s}} = 13$ TeV. *Phys. Rev. D*, 100(5):052011, 2019. doi: 10.1103/PhysRevD.100.052011.

[139] Julian Katz-Samuels, Gilles Blanchard, and Clayton Scott. Decontamination of mutual contamination models. *arXiv preprint arXiv:1710.01167*, 2017.

[140] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models - going beyond svd, 2012.

[141] Cédric Févotte and Jérôme Idier. Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, 23(9):2421–2456, 2011.

[142] David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in neural information processing systems*, pages 1141–1148, 2004.

[143] Andrzej Cichocki, Rafal Zdunek, and Shun-ichi Amari. Csiszar's divergences for non-negative matrix factorization: Family of new algorithms. In *International Conference on Independent Component Analysis and Signal Separation*, pages 32–39. Springer, 2006.

[144] Chris Ding, Xiaofeng He, and Horst D Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM international conference on data mining*, pages 606–610. SIAM, 2005.

[145] Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(Nov):1457–1469, 2004.

[146] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[147] Valentin Leplat, Nicolas Gillis, and Man Shun Ang. Blind audio source separation with minimum-volume beta-divergence nmf, 2019.

[148] Dennis L Sun and Cedric Fevotte. Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6201–6205. IEEE, 2014.

[149] H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492, Berkeley, Calif., 1951. University of California Press. URL https://projecteuclid.org/euclid.bsmsp/1200500249.

[150] Andrzej Cichocki, Rafal Zdunek, and Shun-ichi Amari. New algorithms for non-negative matrix factorization in applications to blind source separation. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V–V. IEEE, 2006.

[151] Mikkel N Schmidt and Rasmus K Olsson. Single-channel speech separation using sparse non-negative matrix factorization. In *Ninth International Conference on Spoken Language Processing*, 2006.

[152] Zhihui Zhu, Xiao Li, Kai Liu, and Qiuwei Li. Dropping symmetry for fast symmetric nonnegative matrix factorization. In *Advances in Neural Information Processing Systems*, pages 5154–5164, 2018.

[153] M Catral, Lixing Han, Michael Neumann, and Robert J Plemmons. On reduced rank nonnegative matrix factorization for symmetric nonnegative matrices. *Linear Algebra and its Applications*, 393:107–126, 2004.

[154] Adrian S Lewis and Jérôme Malick. Alternating projections on manifolds. *Mathematics of Operations Research*, 33(1):216–234, 2008.

[155] Dmitriy Drusvyatskiy, Alexander D Ioffe, and Adrian S Lewis. Transversality and alternating projections for nonconvex sets. *Foundations of Computational Mathematics*, 15(6):1637–1651, 2015.

[156] Brendan O'Donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068, 2016.

[157] Da Kuang and Haesun Park. Fast rank-2 nonnegative matrix factorization for hierarchical document clustering. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 739–747, 2013.

[158] Rundong Du, Da Kuang, Barry Drake, and Haesun Park. Hierarchical community detection via rank-2 symmetric nonnegative matrix factorization. *Computational social networks*, 4(1):7, 2017.

[159] Trapit Bansal, Chiranjib Bhattacharyya, and Ravindran Kannan. A provable svd-based algorithm for learning topics in dominant admixture corpus, 2014.

[160] Yuval Rabani, Leonard J. Schulman, and Chaitanya Swamy. Learning mixtures of arbitrary distributions over large discrete domains. *CoRR*, abs/1212.1527, 2012. URL http://arxiv.org/abs/1212.1527.

[161] Torbjörn Sjöstrand, Stefan Ask, Jesper R. Christiansen, Richard Corke, Nishita Desai, Philip Ilten, Stephen Mrenna, Stefan Prestel, Christine O. Rasmussen, and Peter Z. Skands. An Introduction to PYTHIA 8.2. *Comput. Phys. Commun.*, 191:159–177, 2015. doi: 10.1016/j.cpc.2015.01.024.

[162] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. FastJet User Manual. *Eur. Phys. J. C*, 72:1896, 2012. doi: 10.1140/epjc/s10052-012-1896-2.

[163] Brendan O'Donoghue, Eric Chu, Neal Parikh, and S Boyd. Scs: Splitting conic solver, version 2.0. 2, 2017.

[164] Steven Diamond and Stephen Boyd. Cvxpy: A python-embedded modeling language for convex optimization. *The Journal of Machine Learning Research*, 17(1):2909–2913, 2016.

[165] Gregor Kasieczka, Ben Nachman, and David Shih. R&D Dataset for LHC Olympics 2020 Anomaly Detection Challenge, April 2019. URL https://doi.org/10.5281/zenodo.2629073.

[166] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi. DELPHES 3, A modular framework for fast simulation of a generic collider experiment. *JHEP*, 02:057, 2014. doi: 10.1007/JHEP02(2014)057.

[167] Patrick T Komiske, Eric M Metodiev, and Jesse Thaler. Energy flow polynomials: A complete linear basis for jet substructure. *Journal of High Energy Physics*, 2018(4):13, 2018.