QCD Masterclass Lectures on Jet Physics and Machine Learning

Andrew J. Larkoski

Email: `larkoa@gmail.com`

June 18, 2024

Abstract

These lectures were presented at the 2024 QCD Masterclass in Saint-Jacut-de-la-Mer, France. They introduce and review fundamental theorems and principles of machine learning within the context of collider particle physics, focused on application to jet identification and discrimination. Numerous examples of binary discrimination in jet physics are studied in detail, including $H \to b\bar{b}$ identification in fixed-order perturbation theory, generic one- versus two-prong discrimination, and up versus down quark jet classification. Quark versus gluon jet discrimination is considered in multiple contexts, from using additive, infrared and collinear safe observables, to using hadronic multiplicity, and to including measurements of the jet charge. While many of the results presented here are well known, some novel results are presented, the most prominent being a parametrized expression for the likelihood ratio of quark versus gluon discrimination for jets on which hadronic multiplicity and jet charge are simultaneously measured. End-of-lecture exercises are also provided.

Contents

1 Introduction and Motivation

Machine learning (ML), or extremely high dimensional functional fitting and extrapolation through optimization of an objective function, is exploding as a research discipline in its own right, as well as through its application to particle physics. Within particle physics, machine learning has been a relatively mature sub-discipline for nearly 10 years, shortly after the rise of graphical processing

units (GPUs) used for extreme parallelization tasks in the early 2010s, to application of standard deep neural network architectures, like convolutional neural networks (CNNs) or recurrent neural networks (RNNs), to binary discrimination problems that are central to data interpretation in particle collision experiments. Since this time in the mid-2010s, machine learning in particle physics has very literally increased exponentially, from a handful of papers each year, to now well over 1000. As such, I will not attempt to thoroughly cite the particle physics machine learning literature, and instead point to reviews (some of which are now nearly historical documents) [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17] and to the HEP ML Living Review [18] which continually updates relevant references. For a practical introduction to the methods and successes of ML applied to collider physics, there have been a number of community or public data analysis competitions, e.g., Refs. [19, 20, 21, 22, 23, 24]. A number of yearly physics conferences are dedicated to particle physics analyses with machine learning, including BOOST [25], ML4Jets [26], and Hammer & Nails [27], and the large machine learning conferences now typically include a science workshop in which particle physics is well-represented, such as at ICML [28] or NeurIPS [29]. This is already a very incomplete list that will assuredly grow in the future.

Within particle physics, machine learning was first explored and applied within jet substructure, the study of the internal structure of high-energy, collimated streams of particles ubiquitous in modern high-energy collision experiments like the Large Hadron Collider (LHC). A display of a di-jet event from early Run 3 in the ATLAS experiment is shown in Fig. 1. Jets are a signature phenomena in QCD, whose observation validated the quark and parton models in deeply inelastic scattering experiments of the late 1960s [30, 31] and in e^+e^- collisions of the mid-1970s [32], and in the late 1970s, led to the discovery of the gluon [33, 34, 35, 36, 37], correspondingly validating the whole of QCD as the theory of the strong nuclear force.[1] As experimental collision energies increased, the produced jets grew in complexity and representative phenomena, consisting of dozens of particles initiated by quarks, gluons, or a whole zoo of other particles that have been discovered in the decades since the 1970s. The central goal of modern jet substructure is focused on establishing properties of the unmeasureable initiating particle from the properties, correlations, and dynamics of the actually experimentally observable particles.

These lectures then serve the dual goals of introducing jets, jet substructure, and techniques for analyzing them, as well as machine learning focusing on its application to the physics of jets. To the latter goal, my review will be rather distinct from other approaches due almost exclusively to my idiosyncratic view on the utility of machine learning. As I will describe through the sections that follow, we will, for the most part, treat the machine that analyzes jets for our particular task and its goals as a black box, merely controlling the input and interpreting the output. We will learn to think like a machine, to sharpen and formalize our specific goal, guided by a collection of fundamental theorems of machine learning that we review. To the former goal, we will introduce the physics of jets through a number of examples of binary discrimination problems, the problem of optimally separating signal from background in some mixed, unlabeled data sample. We will find, through a textbook application of the scientific method, making simple, justifiable assumptions and hypotheses, and following the theory of QCD, that we can deeply and robustly understand a wide range of results and will make new predictions that we can test.

---

[1]The first reference I am aware of that uses the term "jet" in the QCD sense is Ref. [38] from a study of cosmic ray air showers in 1953. As for the etymology of "jet", I do not know for sure, but a plausible source is from the Jet d'Eau, the spectacular water feature in Lake Geneva, just off the shore from downtown Geneva which itself dates from 1886. If you have more information about the origin of the term "jet", please let me know.
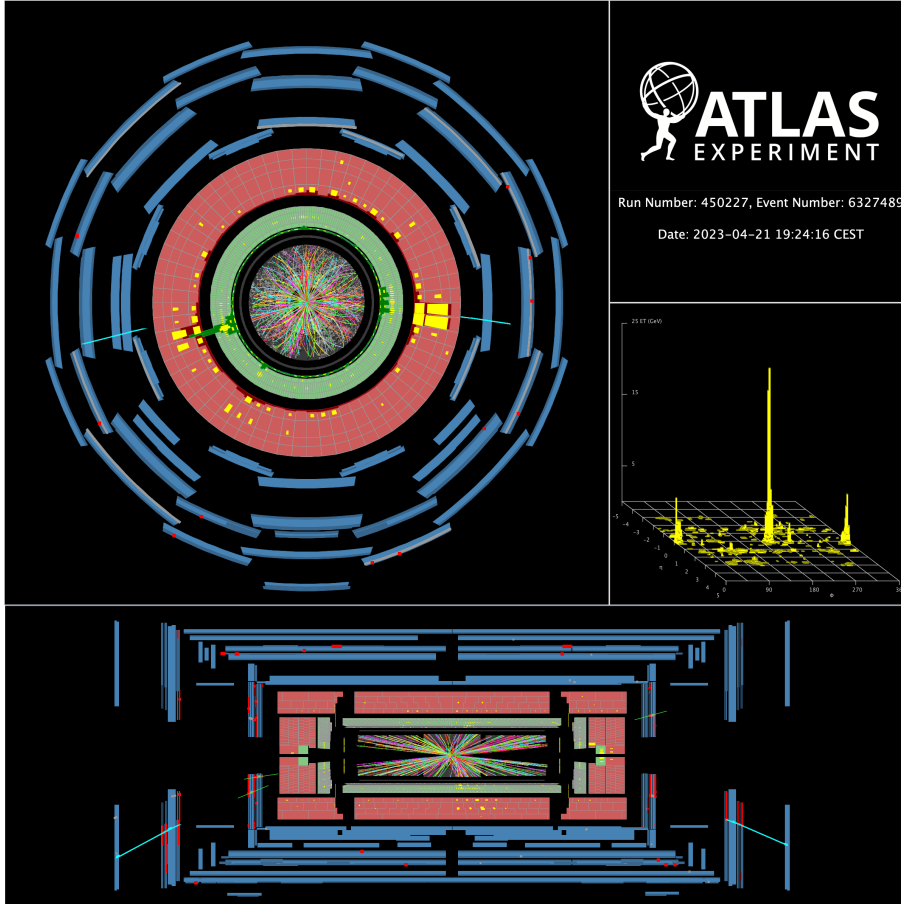
Figure 1: Display of a di-jet event in the ATLAS experiment from early in Run 3 from April 2023. Jets are most obviously identified in the inset on the center right, in the so-called "lego plot", which shows energy deposits in the calorimetry in the unrolled cylindrical detector. Jets are the high-energy, collimated (tall, narrow spikes) in the lego plot.

Nevertheless, these lectures do not exist in a vacuum, and there are numerous other extant review articles that introduce jets and jet substructure from a variety of viewpoints. Several references on jets for theorists that I revisit myself are Refs. [39, 40, 41, 42, 43, 44, 45, 46, 47, 48]. Though now over 30 years old, the "QCD Pink Book" [49] is an absolutely invaluable reference for the history of the topic, a compendium of fundamental results, and correct normalization of important quantities. I have lectured on jets at several previous summer and winter schools, and several of those lectures are on arXiv [50, 51, 52]. My philosophy for both teaching and presenting theoretical physics topics is expounded on in my textbooks on particle physics [53] and quantum mechanics [54]. Additionally, with the rise of machine learning and artificial intelligence, many theoretical physicists have made the transition from academia to industry and have further performed the service of translation of results into a form familiar to physicists. A few reviews of modern machine learning by theoretical physicists for theoretical physicists are Refs. [55, 56, 57].

The outline of these lectures is as follows. In Sec. 2, I will review what I consider the fundamental

Figure 2: Our theorist's impression of a machine learning architecture.

theorems of machine learning, specifically for the application of binary discrimination problems in particle physics. These results are the universal approximation theorem, the central limit theorem, and the Neyman-Pearson lemma. In Sec. 3, I will review fundamental principles from quantum field theory and QCD necessary to make theoretical predictions in jet physics. These results include the master formula for differential cross sections (probability distributions), infrared and collinear safety, collinear and soft factorization, and some remarks about jet algorithms (though we won't worry much about specific algorithms in applications). With this foundation set, the following four sections study four different binary discrimination problems and stitch together the machine learning and quantum field theory results in various contexts. In Sec. 4, we introduce quark versus gluon jet discrimination through a simple picture of scale invariant emissions off of the initiating particle. In Sec. 5, we study binary discrimination within fixed-order perturbation theory with application to identification of $H \to b\bar{b}$ decays, which was in fact the problem that jump-started all of modern jet substructure. In Sec. 6, we return to quark versus gluon discrimination, but this time focusing on particle multiplicity itself and in doing so take our first small steps away from the warmth and comfort of perturbation theory. In Sec. 7 we forgo perturbation theory altogether, which will be necessary for discrimination of up-flavor quark jets from down-flavor jets, but, guided by the solid handrails of statistics, we will nevertheless be able to make many powerful predictions. We summarize in Sec. 8, and present some thoughts on the future of machine learning in particle physics.

2   Some Fundamental Ideas from Machine Learning and Statistics

We start this lecture in earnest by defining machine learning, especially in the context we will use throughout these notes. At its most basic form, our picture of a "machine" is represented in Fig. 2. The machine, the black box, takes input on the left, analyzes it in some way, and returns some output on the right. Here, we will not inquire as to *how* the machine is analyzing the data, or how it is trained, or how many parameters it has, but will rather use this incredibly simple picture to reframe our way of thinking about specific problems in physics to establish their solution. A bit concretely, the more detailed picture on the right of Fig. 2 is what we will focus on specifically. Here, at left, a large number n of data points is input to the machine, each of which live in some enormous-dimensional space (hundreds or thousands of dimensions). The machine analyzes these data and then outputs a function $\mathcal{L}(\vec{x})$ which classifies the known data and extrapolates to unseen data elsewhere in the space.
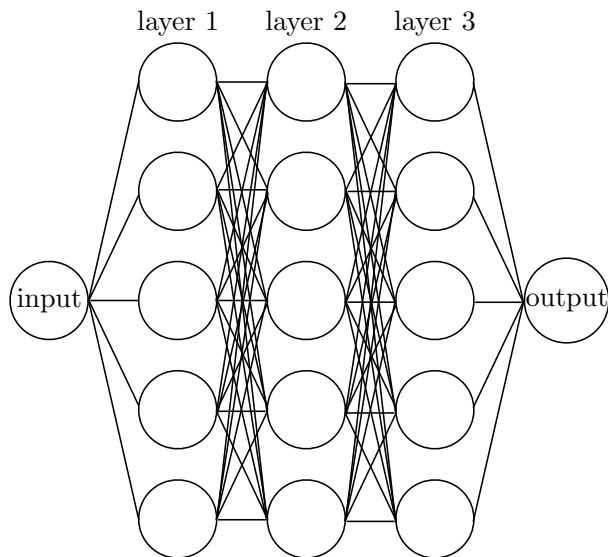
Figure 3: A schematic drawing of a three-layer fully-connected multilayer perceptron architecture. The empty circles represent individual neurons and the lines represent connections between them. The input, at left, is distributed across the first layer of neurons, then passed to the next two "hidden" layers, and then linearly combined to form the output at right.

We will be particularly interested in the situation when the input data is divided into two classes, which we call signal and background, respectively, and the function the machine learns is the optimal observable on the space to discriminate these two classes. This is a very general problem in statistics called **binary discrimination** and a machine's success in solving it depends on two fundamental results. First, the optimal observable for discrimination is in general some arbitrary function of the input data on its high-dimensional domain. Therefore, the machine must be able to approximate this optimal discriminant observable to arbitrary accuracy. Second, the machine needs to know what its target is, it needs to know what exactly is this optimal discrimination observable. Knowing its target, the machine can then optimize its internal parameters, or *learn*, to match the optimal discriminant as closely as possible. Fortunately for us, both of these requirements have been solved, the first referred to as the universal approximation theorem and the second as the Neyman-Pearson lemma, and we will review both in detail below. We will also review the central limit theorem and the three of these results form the basis for everything that follows here.

2.1 Universal Approximation Theorem

To understand the validity of our requirement that a machine can indeed approximate any function of the input to arbitrary accuracy, we need to peek beneath the black box curtain just a bit to understand its internal structure. A modern, feed-forward, fully-connected deep neural network has an internal architecture as modeled by Fig. 3. For historical reasons, this is also called a **multilayer perceptron**, or MLP [58, 59, 60, 61]. This will be our prototypical machine learning architecture and the only that we consider here, but should only be considered representative because there are now many architectures (CNN [62], RNN [63, 64, 65, 66, 67], generative adversarial networks [68],

7

decoder-encoder transformers [69], etc.) that are explored and used depending on the particular application (image recognition, language translation, prompt completion, etc.).[2]

2.1.1 Very Brief Description of Practical Machine Learning

In Fig. 3, the input data are distributed across the first layer of neurons, then that data are processed and passed to the next layer, and continuing, until a linear combination is taken of the processed data to produce the output. Layers after the first layer that directly interacts with the input are called "hidden" because they are hidden from the input. Each layer of the MLP processes the data from the previous layer by an affine transformation composed with a non-linear activation function σ. That is, going from layer ℓ to layer $\ell + 1$, the neurons act on the response from layer ℓ as:

$$\vec{x}_{\ell+1} = \sigma \left(\mathbb{W}_\ell \vec{x}_\ell + \vec{b}_\ell \right) . \tag{2.1}$$

Here, $\mathbb{W}_\ell$ is called the weight matrix, $\vec{b}_\ell$ is a bias vector, and $\vec{x}_\ell$ is the vector of all neuron outputs at layer ℓ. For example, if layer ℓ has n_ℓ neurons, then $\vec{x}_\ell$ is n_ℓ dimensional. The non-linear activation function σ acts on all elements of its vector argument individually.

What makes machine learning fascinating and very distinct from traditional programming is that the weights $\mathbb{W}_\ell$ and the biases $\vec{b}_\ell$ are not preprogrammed; they are learned through minimization of an objective function $\mathcal{L}$ that compares input to output and is a function of all of the weights and biases. With appropriate normalization, $\mathcal{L} = 0$ means that the weights and biases have been learned to perfectly align the output with the input. Learning is then accomplished through **gradient descent** of the objective function. Elements of the weights and biases are considered as parameters of the neural network and are initialized according to some expected probability distribution. This is typically chosen to be Gaussian, and we will provide a physics interpretation of this in a bit. Then, with initialized parameters set, the objective function is evaluated and the parameters are updated by flowing in the direction opposite to the gradient of the objective function, and then the objective function is evaluated again with the updated parameters. Each complete pass through the dataset updating network parameters is called an **epoch**, and a weight parameter/element $\mathbb{W}_{\ell,ij}$ is updated as

$$\mathbb{W}_{\ell,ij}^{(m+1)} = \mathbb{W}_{\ell,ij}^{(m)} - \eta \nabla_{\mathbb{W}_{\ell,ij}} \mathcal{L}(\{\mathbb{W}^{(m)}, \vec{b}^{(m)}\}) . \tag{2.2}$$

Here, the superscript (m) denotes the parameters at the mth epoch, $\nabla_{\mathbb{W}_{\ell,ij}}$ is the derivative with respect to the parameter of interest, and η is called the **learning rate**, a parameter that controls the rate of descent down the gradient. If you pass through the network enough epochs, the value of the objective function will stop changing and you have therefore learned the appropriate function of the input data.

There are a few comments I want to make before making the statement of the universal approximation theorem. First, this naive gradient descent is never used in modern applications because

---

[2]Very recently, there has been a proposal for a significantly different deep learning architecture from the MLP/activation function/universal approximation theorem paradigm [70]. This network is based on the result of the Kolomogorov-Arnold theorem [71, 72, 73, 74, 75] which states that any n-dimensional multivariate function can be expressed as a linear combination of $2n+1$ continuous functions whose arguments are linear combinations of n single-variable functions. However, the simplicity, utility, or relevance of the Kolmorgorov-Arnold theorem as opposed to the universal approximation theorem to machine learning architecture is highly disputed [76, 77, 78, 79, 80, 81, 82, 83, 84] so much more work will need to be done to determine if it is indeed a game-changer.

it is much, much too costly to evaluate if your data set and network parameters are large (e.g., billions or even trillions of elements). Instead, methods like **stochastic gradient descent** [85, 86], in which smaller **batches** are randomly or stochastically selected during every epoch to train the network, are used. Further, the gradient itself can be problematic to evaluate because the objective function on parameter space can be extremely non-smooth, so many methods for smoothing the gradient have been introduced like **momentum** [87] or **Adam** [88]. Because of the recursive composition nature of the neurons on the network, the gradient can be evaluated through application of the Leibniz chain rule, which is very efficient if the derivative of the non-linear activation function σ is simple. A particularly efficient way to evaluate the chain rule is through **backpropagation**, which simply means evaluating the derivatives of functions in composition backward through the network. In terms of function composition, this means evaluating derivatives from the outside (the most recent evaluation) in (towards the earliest evaluation). This ensures that derivatives of parameters early in the layers can be reused efficiently.

2.1.2 Statement of the Universal Approximation Theorem

With this setup for our neural network, we want to establish the conditions by which the output can indeed be an arbitrary function of the input. The result is called the **universal approximation theorem** whose original proofs made rather strong, non-physical assumptions on the network [89, 90, 91]. Concretely, the original proofs assumed that the non-linear activation function was a sigmoid, with

$$\sigma(x) = \frac{1}{1 + e^{-x}} \, , \tag{2.3}$$

which was a popular activation function at the time (late 1980s, early 1990s). Cybenko's proof assumed that the network consisted of a single layer, but was infinitely wide. In this case, the statement of the universal approximation theorem is that such single-layer, infinitely-wide networks could approximate any function to arbitrary accuracy; or, that arbitrary linear combinations of sigmoids could approximate any function.

While we won't present a proof of it, let's be a bit more specific and state the universal approximation precisely (at least in the form of Cybenko's proof). Let your data be represented by a vector $\vec{x}$. We can then construct the function $G(\vec{x})$ from a linear combination of sigmoids as

$$G(\vec{x}) = \sum_{i=1}^{N} c_i \, \sigma \left(\mathbb{W}_i \vec{x} + \vec{b}_i \right) \, , \tag{2.4}$$

where $\mathbb{W}_i$ is a weight matrix, $\vec{b}_i$ is a bias vector, and c_i is a real-valued coefficient. The number of terms in the sum, N, is merely assumed to be finite, $N < \infty$. Then, with appropriate choice of the weights, biases, and coefficients, the function $G(\vec{x})$ can approximate any function $f(\vec{x})$ to arbitrary accuracy. Specifically, for any $\epsilon > 0$, there exist weights, biases, and coefficients such that

$$|f(\vec{x}) - G(\vec{x})| < \epsilon \, . \tag{2.5}$$

No relationship between the number of terms N in the sum and the desired accuracy ϵ was established in the early proofs, and in Cybenko's paper, he specifically mentions that for any reasonable accuracy ϵ, N is likely enormous. Indeed, for *arbitrary* accuracy, the number of terms in the sum must be infinite.

Today, sigmoids are basically never used anymore as activation functions and the machine learning community has explored a whole zoo of possible functions. When using gradient descent for parameter/weight optimization, a problem with the sigmoid is that taking the gradient of many composed sigmoids uses the chain rule, that is, the gradient is sensitive to the derivatives at every level of the network. Derivatives of the sigmoid are bounded on $\sigma'(x) \in [0,1]$, and away from its "turn on" at $x = 0$, rapidly vanish. This can lead to catastrophic loss of correlation between neurons even a few layers apart because just a few products of sigmoid derivatives can quickly get very small and is in fact exponential in the number of layers. This problem with the sigmoid or other activation functions with small derivatives is called the **vanishing gradient problem**.

There are many solutions to this problem but perhaps the easiest for these lectures is simply changing the non-linear function to something with a constant derivative almost anywhere. The most popular activation function used now is the rectified linear unit, or **ReLU**, which has the functional form [92]

$$\mathrm{ReLU}(x) = x\,\Theta(x)\,. \tag{2.6}$$

This function is 0 when $x < 0$ and linear with slope 1 when $x > 0$. Thus, derivatives evaluated in backpropagation are simply 1 if the neuron is activated or "fires" (i.e., when $x > 0$). Because of its ubiquity, the universal approximation theorem has been extended to include ReLU and other activation functions, e.g., Refs. [93, 94, 95, 96, 97, 98].

2.1.3 Our Physicist's Interpretation of the Universal Approximation Theorem

Unfortunately, the universal approximation theorem is of rather minimal relevance for practical machine learning. At least as originally stated, it requires an infinitely wide network (clearly unphysical!) and in this limit, for any finite input dataset, we can always simply fix the parameters of the network to memorize the dataset. So, indeed, it can output an arbitrary function of the input, but no active learning was taking place. However, finite networks cannot simply memorize any input dataset, and so learning or training of parameters must take place and the rate of convergence of learning the objective function, residual error given network architecture, etc., are not in the purview of the universal approximation theorem. Further, the proofs are typically non-constructive: they tell you that a sufficiently large and powerful network can approximate any function, but not how to do it.

So, what good is the universal approximation theorem for us? Well, as physicists, the precise architecture and parameters of a practical network are not directly our focus; we want to learn (as humans!) the objective function for our specific task. That is, given any possible machine we can determine optimal performance, given the laws of physics. Additionally, with this understanding, the universal approximation theorem allows us to represent our data in whatever form is convenient for calculation or human interpretation more broadly. Because a sufficiently powerful machine can learn *any* function of the input data, as long as the representation of the input data always completely spans its domain, a sufficiently powerful machine will always learn the same output or objective function. This is represented in Fig. 4. At left, the data are in the form of a vector $\vec{x}$ and the machine learns function $g(\vec{x})$. At right, we have transformed the initial data $\vec{x}$ by some set of functions $\{f_i(\vec{x})\}_i$ and as long as those functions are non-degenerate, the machine will learn the exact same function $g(\vec{x})$.

Thus, as physicists, we will always imagine working with a machine that has infinite and representative input data (the machine does no extrapolation) and is infinitely powerful (learns the

$$x_1, x_2, x_3, \dots, x_n \longrightarrow g(\vec{x}) \qquad f_1(\vec{x}), f_2(\vec{x}), f_3(\vec{x}), \dots, f_n(\vec{x}) \longrightarrow g(\vec{x})$$

Figure 4: The consequence of the universal approximation theorem: the output $g(\vec{x})$ will always be learned by a sufficiently powerful machine as long as the input information spans the data space of interest.

objective function perfectly). As such, we can choose to represent the input data in whatever form we want as long as it still completely spans the data domain in the sense that the original form of the data can be evaluated losslessly (i.e., the functions $\{f_i(\vec{x})\}_i$ in Fig. 4 are injective). By considering what we can actually calculate and predict within quantum field theory and QCD, we will be lead to particular, preferred representations of the input data, but again, that is merely for convenience and our bias as theorists that we like to calculate. That is, we want to use the theory of the physics we are studying to determine the absolute best possible performance that any such machine could produce.

With that, we can safely close the lid on the machine and for the rest of these lectures treat it as a black box.

2.2 Neyman-Pearson Lemma and Binary Discrimination

Now that our machine can output an arbitrary function of the input, we need to actually determine what output we want! As mentioned earlier, throughout these lectures, we will focus on binary discrimination problems in which we would like to determine the observable (i.e., function of the inputs) that optimally separates these two classes. This is a very general problem especially in particle physics where, given some data event from the LHC say, we would like to classify it as "interesting" or signal, or "not interesting" or background. In a couple of lectures we will introduce perhaps the historically most important such classification problem in jet substructure, namely, distinguishing the dominant decay mode of the Higgs boson, $H \to b\bar{b}$, from the QCD background process of gluon splitting to bottom quarks, $g \to b\bar{b}$.

Fortunately for us, the optimal discrimination observable has been identified through the **Neyman-Pearson lemma** as the **likelihood ratio**, or the ratio of the probability distributions of signal and background. In the explicit examples we consider in the following lectures, our goal will then be to first determine the signal and background probability distributions on an appropriate data space and then from those distributions, determine the likelihood ratio. However, simply given the likelihood ratio, this only tells us it is the optimal discriminant given the input data representation, but the actual performance for binary discrimination will depend on the specific functional relationships of the signal and background distributions. Here, we will also introduce the **receiver operating characteristic** as a very useful representation of the discrimination power, once we have identified the likelihood ratio.

2.2.1 Statement and Proof of the Neyman-Pearson Lemma

We will start, however, with simply the statement of the Neyman-Pearson lemma [99] and provide its simple proof. We assume that we measure some variables x and that on the domain of x, signal events are distributed according to the distribution $p_s(x)$ and background events are distributed as $p_b(x)$. We will often use this subscript notation to denote event labels, but a more explicit notation is through conditional probabilities, where

$$p_s(x) = p(x|s)\,, \qquad\qquad p_b(x) = p(x|b)\,. \qquad (2.7)$$

That is, $p(x|s)$ is the distribution of observable variables x given that events are drawn from the signal. Additionally, observables x here may be vectors, or some other data representation, but we will simply write x as the specific form of the data is not relevant.

Given this setup, we now define the constraints in the Neyman-Pearson lemma. We would like to determine the region on x space that maximizes the signal, for a fixed amount of background contamination. The region can be specified by the requirement that $\mathcal{L} > \Lambda(x)$, where $\Lambda(x)$ is some function of the data x whose form we will optimize for. The background contamination or **false positive rate** is simply the probability that the background events lie in the region of interest, which is

$$\alpha = \int dx\, p_b(x)\, \Theta\left(\mathcal{L} - \Lambda(x)\right)\,, \qquad (2.8)$$

and by assumption, the rate α is fixed. The **true positive rate** that we want to maximize is equivalent to maximizing the difference

$$\int dx\, (p_s(x) - p_b(x))\, \Theta\left(\mathcal{L} - \Lambda(x)\right)\,, \qquad (2.9)$$

because the false positive rate is fixed.

This is all the setup we need. This is now a simple constrained optimization problem that can be solved by Lagrange multipliers. We introduce the Lagrange multiplier λ that multiplies the constraint in the function we wish to maximize as

$$\int dx\, (p_s(x) - \lambda p_b(x))\, \Theta\left(\mathcal{L} - \Lambda(x)\right)\,. \qquad (2.10)$$

Next, we want to optimize the region $\Lambda(x)$, so we take a variational derivative with respect to $\Lambda(x)$ and set it to 0:

$$\frac{\delta}{\delta\Lambda(x)} \int dx\, (p_s(x) - \lambda p_b(x))\, \Theta\left(\mathcal{L} - \Lambda(x)\right) = -\int dx\, (p_s(x) - \lambda p_b(x))\, \delta\left(\mathcal{L} - \Lambda(x)\right) = 0\,. \quad (2.11)$$

Now, we immediately see that this is trivially satisfied, the integrand itself vanishes, if we set the Lagrange multiplier $\lambda = \mathcal{L}$ and the region of interest defined by

$$\Lambda(x) = \frac{p_s(x)}{p_b(x)}\,, \qquad (2.12)$$

which is the likelihood ratio. Thus, the likelihood ratio is the optimal binary discriminant that maximizes signal given a fixed false positive rate. Actually, we can weaken the statement of the Neyman-Pearson lemma a bit because the original optimization problem, Eq. (2.9), is unchanged if $\Lambda(x)$ is any monotonic function of the likelihood ratio. So, the optimal discrimination observable is monotonic in the likelihood, by the Neyman-Pearson lemma.

2.2.2 ROC Curves and Representations of Discrimination Power

Now that we know the optimal discrimination observable is the likelihood ratio, we would like to quantify its discrimination power. One approach, given the discussion above, would be to plot the true positive rate $\Sigma_s(\alpha)$ as a function of the false positive rate α defined as

$$\Sigma_s(\alpha) = \int dx\, p_s(x)\, \Theta\left(\mathcal{L} - \frac{p_s(x)}{p_b(x)}\right), \qquad \alpha = \int dx\, p_b(x)\, \Theta\left(\mathcal{L} - \frac{p_s(x)}{p_b(x)}\right). \tag{2.13}$$

By the Neyman-Pearson lemma, for a fixed α, the likelihood ratio maximizes the true positive rate, $\Sigma_s(\alpha)$. Note that the dependence of the true positive rate on the false positive rate α is implicit, through the value of $\mathcal{L}$, so let's work to make it explicit. To do this, let's massage the false positive rate as

$$\int dx\, p_b(x)\, \Theta\left(\mathcal{L} - \frac{p_s(x)}{p_b(x)}\right) = \int_0^{\mathcal{L}} d\mathcal{L}' \int dx\, p_b(x)\, \delta\left(\mathcal{L}' - \frac{p_s(x)}{p_b(x)}\right) \tag{2.14}$$

$$= \int_0^{\mathcal{L}} d\mathcal{L}'\, p_b(\mathcal{L}') = \Sigma_b(\mathcal{L}),$$

where $\Sigma_b(\mathcal{L})$ is the cumulative probability distribution of the likelihood on the background sample. Because probability density $p_b(x)$ is everywhere non-negative, the cumulative distribution $\Sigma_b(\mathcal{L})$ is monotonic, and so its inverse is well-defined. Then, the value of the likelihood ratio at fixed false positive rate α is

$$\mathcal{L} = \Sigma_b^{-1}(\alpha). \tag{2.15}$$

From similar manipulations of the true positive rate, it is also a cumulative probability distribution and we give the true positive rate as a function of false positive rate a name, where

$$\text{ROC}(\alpha) \equiv \Sigma_s\left(\Sigma_b^{-1}(\alpha)\right), \tag{2.16}$$

which is called the **receiver operating characteristic curve** or ROC curve. It ranges from $\text{ROC}(x) \in [0,1]$ for $x \in [0,1]$ and is bounded from above by the line $\text{ROC}(x) = x$. In fact, if $\text{ROC}(x) = x$ then the true positive and false positive rates are equal; that is, the observable completely randomly selects events as signal or background.

Because the ROC curve is the minimum false positive rate for a given true positive rate, its integral is a useful measure of discrimination, as well. The integral of the ROC curve, or the area under the curve (AUC) is then

$$\text{AUC} = \int_0^1 dx\, \Sigma_s\left(\Sigma_b^{-1}(x)\right) = \int d\mathcal{L}_b\, p_b(\mathcal{L}_b)\, \Sigma_s(\mathcal{L}_b) \tag{2.17}$$

$$= \int d\mathcal{L}_s\, d\mathcal{L}_b\, p_s(\mathcal{L}_s)\, p_b(\mathcal{L}_b)\, \Theta(\mathcal{L}_b - \mathcal{L}_s).$$

By the definition of the likelihood ratio,

$$\mathcal{L} = \frac{p_s(x)}{p_b(x)}, \tag{2.18}$$

the likelihood is large where the signal distribution is large, and small where the background distribution is large. Thus, the AUC is a measure of the "mis-ordering" of events in the likelihood. Perfect discrimination, for which all signal events lie at larger values of the likelihood than all background events, has an AUC of 0, while completely random discrimination, for which signal and background events are uniformly mixed in the likelihood, has an AUC of 1/2. Therefore, a standard objective function for binary discrimination that is often used in machine learning is the AUC.

It's worth diving a little bit deeper into the relationship between the distributions of the likelihood ratio $\mathcal{L}$ on signal and background distributions. First, from the definition of the likelihood, we have

$$\mathcal{L} = \frac{p_s(\mathcal{L})}{p_b(\mathcal{L})}, \qquad (2.19)$$

so, obviously,

$$p_s(\mathcal{L}) = \mathcal{L}\, p_b(\mathcal{L})\,. \qquad (2.20)$$

Then, because both signal and background distributions of the likelihood ratio are normalized, this requires that the mean of the background distribution is unity:

$$1 = \int_0^\infty d\mathcal{L}\, p_b(\mathcal{L}) = \int_0^\infty d\mathcal{L}\, p_s(\mathcal{L}) = \int_0^\infty d\mathcal{L}\, \mathcal{L}\, p_b(\mathcal{L})\,. \qquad (2.21)$$

The AUC can also be expressed exclusively in terms of the background distribution of the likelihood. First, from Eq. (2.17), we can also write

$$\text{AUC} = 1 - \int_0^\infty d\mathcal{L}\, p_s(\mathcal{L})\, \Sigma_b(\mathcal{L}) = 1 - \int_0^\infty d\mathcal{L}\, \mathcal{L}\, p_b(\mathcal{L})\, \Sigma_b(\mathcal{L})\,. \qquad (2.22)$$

Using integration-by-parts, this expression exclusively in terms of the background distribution can be expressed as

$$\text{AUC} = 1 - \frac{1}{2}\int_0^\infty d\mathcal{L}\left[1 - \Sigma_b(\mathcal{L})^2\right]\,. \qquad (2.23)$$

So, if you have an independent bound on the functional form of the background cumulative distribution of the likelihood, this can provide a useful estimate of the AUC.

For its intuitive definition, the ROC curve is rather unintuitively named, and one might wonder how such a seemingly useful statistical measure of discrimination power has such a clunky title. The true positive versus false positive rate plot was developed in World War II for quantifying the efficacy ("operating characteristic") of receivers (the signal processing units of a radar, which was at the time often a person) to identify whether a blip was a plane or not a plane. The information contained in the ROC curve was formalized shortly thereafter, e.g., Refs. [100, 101], which was also shortly after Shannon's theory of information [102]. While the ROC is still widely used in radar applications today, its utility has extended far beyond, but the original name has stuck.

2.2.3 Robust Lower Bound of the AUC

With the expression for the AUC in Eq. (2.23), we can rather simply determine a robust lower bound, as determined by the minimum $\mathcal{L}_{\min} \geq 0$ and maximum $\mathcal{L}_{\max} < \infty$ values of the likelihood ratio. Assuming that the minimum and maximum are not 0 and infinite, respectively, the expression for the AUC can be expressed as an integral exclusively over the range $[\mathcal{L}_{\min}, \mathcal{L}_{\max}]$ as

$$\text{AUC} = 1 + \frac{1}{2} \int_{\mathcal{L}_{\min}}^{\mathcal{L}_{\max}} d\mathcal{L} \, \Sigma_b(\mathcal{L})^2 - \frac{\mathcal{L}_{\max}}{2} \,. \tag{2.24}$$

Additionally, the fact that the mean value of the likelihood of the background distribution is unity constrains the integral of the cumulative distribution, where

$$\int_{\mathcal{L}_{\min}}^{\mathcal{L}_{\max}} d\mathcal{L} \, \Sigma_b(\mathcal{L}) = \mathcal{L} \, \Sigma_b(\mathcal{L})|_{\mathcal{L}_{\min}}^{\mathcal{L}_{\max}} - \int_{\mathcal{L}_{\min}}^{\mathcal{L}_{\max}} d\mathcal{L} \, \mathcal{L} \, p_b(\mathcal{L}) \tag{2.25}$$

$$= \mathcal{L}_{\max} - 1 \,.$$

That is, we would like to determine the cumulative distribution $\Sigma_b(\mathcal{L})$ that minimizes the AUC with fixed integral on the domain $\mathcal{L} \in [\mathcal{L}_{\min}, \mathcal{L}_{\max}]$. As a constrained minimization problem, we can again use Lagrange multipliers to solve it.

Let's introduce the "action" as the sum of the value of the AUC and the Lagrange multiplier λ times the integral of the cumulative distribution:

$$S[\Sigma_b, \lambda] = 1 + \frac{1}{2} \int_{\mathcal{L}_{\min}}^{\mathcal{L}_{\max}} d\mathcal{L} \left[\Sigma_b(\mathcal{L})^2 + 2\lambda \, \Sigma_b(\mathcal{L}) \right] - \frac{\mathcal{L}_{\max}}{2} \,. \tag{2.26}$$

Extremizing the variational derivative with respect to Σ_b then constrains the Lagrange multiplier as

$$\frac{\delta S[\Sigma_b, \lambda]}{\delta \Sigma_b} = \int_{\mathcal{L}_{\min}}^{\mathcal{L}_{\max}} d\mathcal{L} \, (\Sigma_b(\mathcal{L}) + \lambda) = \mathcal{L}_{\max} - 1 + \lambda \, (\mathcal{L}_{\max} - \mathcal{L}_{\min}) = 0 \,. \tag{2.27}$$

The integrand of this expression vanishes identically if

$$\Sigma_b^{(\min)}(\mathcal{L}) = -\lambda = \frac{\mathcal{L}_{\max} - 1}{\mathcal{L}_{\max} - \mathcal{L}_{\min}} \,, \tag{2.28}$$

which corresponds to the cumulative distribution that minimizes the AUC with the constraints. This constant cumulative distribution corresponds to the background probability distribution of

$$p_b^{(\min)}(\mathcal{L}) = \frac{\mathcal{L}_{\max} - 1}{\mathcal{L}_{\max} - \mathcal{L}_{\min}} \, \delta \, (\mathcal{L} - \mathcal{L}_{\min}) + \frac{1 - \mathcal{L}_{\min}}{\mathcal{L}_{\max} - \mathcal{L}_{\min}} \, \delta \, (\mathcal{L} - \mathcal{L}_{\max}) \,, \tag{2.29}$$

which is normalized and has unit mean.

With this expression for the minimal background cumulative distribution, its AUC can be calculated that forms a lower bound for any AUC given that minimum and maximum likelihood:

$$\text{AUC} \geq \frac{1 - 2\mathcal{L}_{\min} + \mathcal{L}_{\max}\mathcal{L}_{\min}}{2(\mathcal{L}_{\max} - \mathcal{L}_{\min})} \,. \tag{2.30}$$
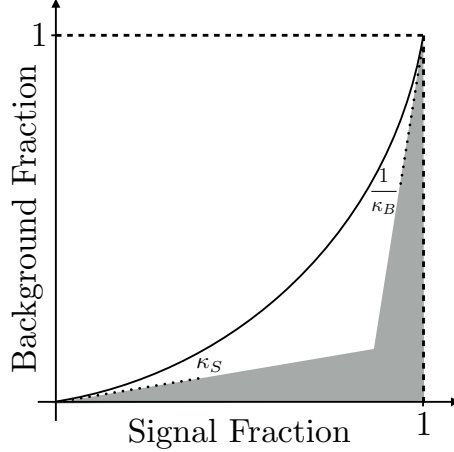
Figure 5: Illustration of a ROC curve and a robust lower bound of the AUC as defined by the area of the gray quadrilateral defined by reducibility factors $\kappa_S = 1/\mathcal{L}_{\max}$ and $\kappa_B = \mathcal{L}_{\min}$. Figure from Ref. [103].

This bound on the AUC was first derived (to my knowledge) in Ref. [103] graphically, from the plot of the ROC curve itself. This is illustrated in Fig. 5. The slope of the ROC curve at its endpoints is set by the minimum and maximum values of the likelihood ratio, and from these slopes a minimal quadrilateral can be formed, as the derivative of the ROC curve is the value of the likelihood at a given background quantile x:

$$\frac{d}{dx}\text{ROC}(x) = \frac{d}{dx}\Sigma_s\left(\Sigma_b^{-1}(x)\right) = \frac{p_s\left(\Sigma_b^{-1}(x)\right)}{p_b\left(\Sigma_b^{-1}(x)\right)} = \Sigma_b^{-1}(x)\,. \tag{2.31}$$

By monotonicity of the likelihood, any ROC curve with those bounds on the likelihood must lie above this quadrilateral. The area of this quadrilateral is precisely the expression of Eq. (2.30). These extremal values of the likelihood ratio are therefore useful measures of discrimination power in their own right, and have come to be called **reducibility factors**, or maximal purity [104, 105, 106]. The minimum and maximum values of the likelihood ratio, by definition determine the maximal possible purity of signal or background samples that can be attained anywhere in the space.

2.2.4 Aside: Principle of Maximum Likelihood

While I promised we wouldn't dive under the hood of our black box neural net again, we'll just take the tiniest of peeks into a useful objective function for binary discrimination. In binary discrimination, events or instances of the data are labeled with a 0 or 1 to represent signal or background, respectively. Then, using the non-label information of the data, the machine learns the correlation with labels and correspondingly outputs a probability that an event is signal, for instance. To learn these probabilities, we want the machine to optimize an appropriate objective function that connects the labels with the probability distributions of the data. Such an objective function that accomplishes this is

$$H(y,p) \equiv - \sum_{i\in\text{data}} \left[y_i\log p_i + (1-y_i)\log(1-p_i)\right]\,, \tag{2.32}$$

16

which is called the **cross entropy**, where y_i is the event label (0 or 1) and p_i is the probability output that the machine assigns to the data. The cross entropy is non-negative (for the same reasons that Boltzmann or Shannon entropy is non-negative), and so we want to minimize it as an objective function. Note that $H(y, p) = 0$ means that the machine perfectly identifies data labels ($y_i = 0$ means $p_i = 0$ and $y_i = 1$ means $p_i = 1$), along with the limiting relationship that $0 \log 0 = 0$. So minimization of the cross entropy means that the machine performs as best as possible label probability assignments.

The cross entropy is further monotonically related to the likelihood that the machine assigns event probabilities correctly, as it is the negative logarithm of the likelihood. Thus, minimization of the cross entropy is equivalent to maximization of the likelihood, and this general feature of objective functions is called the **principle of maximum likelihood** [107, 108]. The cross entropy is further monotonically related to the **Kullback-Liebler divergence** [109], which, for two probability distributions q and p on random variables x, is defined as

$$D_{\mathrm{KL}}(p||q) = \int dx\, p(x)\, \log \frac{p(x)}{q(x)}\,. \tag{2.33}$$

For continuous distributions, the Kullback-Liebler divergence is often used as the objective function for optimization and its interpretation is that it is the information gained by using the distribution $p(x)$ than the current $q(x)$. It is also the expected value of the logarithm of the likelihood ratio on the distribution $p(x)$.

So, what is this probability that the machine assigns to an event it analyzes? We have discussed the conditional probability $p(\mathcal{L}|s)$, which is the distribution of the likelihood given a signal event, for instance. What we want instead is $p(s|\mathcal{L})$; that is, the probability that an event is signal, given the value of the likelihood. These two probabilities are related by **Bayes's theorem** [110], where

$$p(s|\mathcal{L}) = \frac{p(\mathcal{L}|s)p(s)}{p(\mathcal{L})}\,. \tag{2.34}$$

The probability of the likelihood itself, $p(\mathcal{L})$, can be evaluated by summing over all possible event classes; namely,

$$p(\mathcal{L}) = p(\mathcal{L}|s)p(s) + p(\mathcal{L}|b)p(b)\,. \tag{2.35}$$

Now, $p(s)$ ($p(b)$), for example, is the probability that any single event in some mixed ensemble is signal (background). Note that, because there are only two classes by fiat, $p(b) = 1 - p(s)$. Now, we can plug this into Bayes's theorem and find the result

$$p(s|\mathcal{L}) = \frac{p(\mathcal{L}|s)p(s)}{p(\mathcal{L}|s)p(s) + p(\mathcal{L}|b)p(b)} = \frac{\mathcal{L}\,p(s)}{\mathcal{L}\,p(s) + p(b)}\,. \tag{2.36}$$

On the right, we have used the definition of the likelihood where

$$p(\mathcal{L}|s) = \mathcal{L}\,p(\mathcal{L}|b)\,. \tag{2.37}$$

We note that this probability $p(s|\mathcal{L})$ is itself monotonic in the likelihood, as well.

2.3 Central Limit Theorem

With the universal approximation theorem and the Neyman-Pearson lemma, we now know what objective function to use for binary discrimination and that a sufficiently powerful machine will always be able to learn it. In this section, we provide one more fundamental theorem of machine learning, namely, the **central limit theorem** [111, 112, 113, 114], one of the earliest and most important results in all of probability and statistics. The central limit theorem will appear again and again as the starting point or first approximation for limiting dynamics in jets in the examples we consider, and further, the central limit theorem provides a rather interesting physical interpretation for an infinitely wide neural network.

2.3.1 Statement and Proof of the Central Limit Theorem

The statement and proof of the central limit theorem is as follows. Let the random variables $\{x_i\}_i$ each be independent and identically distributed ("i.i.d."). "Identically distributed" means that the value of each one of the x_is is drawn from the same probability distribution $p(x_i)$ and "independent" means that the value of x_i has no effect on the value of x_j that is drawn from the distribution. With this setup, we then consider the sum of N of these i.i.d. random variables and call it X:

$$X = \sum_{i=1}^{N} x_i \,. \tag{2.38}$$

Our goal will be to determine the distribution of X in the limit that the number of terms in the sum N gets large, $N \to \infty$.

To prove the central limit theorem, it is useful to rescale and subtract the mean values, introducing a new variable $\tilde{X}$, where

$$\tilde{X} = \sum_{i=1}^{N} \frac{x_i - \mu}{\sqrt{N}} \,. \tag{2.39}$$

Here, μ is the mean value of the x random variables,

$$\mu = \langle x \rangle = \int dx \, x \, p(x) \,. \tag{2.40}$$

Note that the mean of $\tilde{X}$ is 0:

$$\langle \tilde{X} \rangle = \int \prod_{i=1}^{N} [dx_i \, p(x_i)] \sum_{j=1}^{N} \frac{x_j - \mu}{\sqrt{N}} = 0 \,. \tag{2.41}$$

Its second moment is therefore also its variance, which we can calculate to be

$$\langle \tilde{X}^2 \rangle = \sigma_{\tilde{X}}^2 = \int \prod_{i=1}^{N} [dx_i \, p(x_i)] \left(\sum_{j=1}^{N} \frac{x_j - \mu}{\sqrt{N}} \right)^2 \tag{2.42}$$

$$= \int \prod_{i=1}^{N} [dx_i \, p(x_i)] \left(\sum_{j=1}^{N} \frac{(x_j - \mu)^2}{N} + 2 \sum_{j<k=1}^{N} \frac{(x_j - \mu)(x_k - \mu)}{N} \right)$$

18

$$= \sigma_x^2$$

On the second line, we have simply expanded out the square and noted that the cross term vanishes because x_i and x_j are independent, for $i \neq j$. Therefore, the variance of $\tilde{X}$ is the same as the variance of x itself, where

$$\sigma_x^2 = \int dx \, (x - \mu)^2 \, p(x) \,. \tag{2.43}$$

Now, we move to calculating the probability distribution of $\tilde{X}$, by integrating over all of the x_i variables with the constraint to evaluate $\tilde{X}$ at its defined value:

$$p(\tilde{X}) = \int \prod_{i=1}^{N} [dx_i \, p(x_i)] \, \delta \left(\tilde{X} - \sum_{i=1}^{N} \frac{x_i - \mu}{\sqrt{N}} \right) \,. \tag{2.44}$$

Now, we could work directly with this distribution and massage it in the large-N limit, but note that because of the structure of the δ-function's argument, this integral is effectively a high-dimensional convolution, or, an integral over a product of factors where the sum of arguments is fixed. As a standard trick to deal with convolutions, we will Laplace transform this integral, as the Laplace transform of a convolution is simply a product of the Laplace transformed integrand factors. In statistics, the Laplace transform of a probability distribution is its **cumulant generating function**, by which repeated differentiation produces the distribution's cumulants. In a more physics language, cumulants are a distribution's connected correlation functions, so the cumulant generating function is like evaluating the path integral with a field source, and then repeated variational differentiation with respect to the source produces the connected correlation functions.

Now, explicitly evaluating the Laplace transform, we find

$$\int d\tilde{X} \, e^{t\tilde{X}} \, p(\tilde{X}) = \int d\tilde{X} \int \prod_{i=1}^{N} [dx_i \, p(x_i)] \, e^{t\tilde{X}} \, \delta \left(\tilde{X} - \sum_{i=1}^{N} \frac{x_i - \mu}{\sqrt{N}} \right) \tag{2.45}$$

$$= \left[\int dx \, p(x) \, e^{t \frac{x - \mu}{\sqrt{N}}} \right]^N = \left(1 + \frac{t^2}{2} \frac{\sigma_x^2}{N} + \mathcal{O}(N^{-3/2}) \right)^N$$

$$= e^{\frac{t^2}{2} \sigma_x^2} \,, \text{ as } N \to \infty$$

On the second line, we have expanded the Laplace transform of the distribution $p(x)$ in powers of the parameter t, and what is particularly interesting is that beyond order t^2, terms in the expansion are suppressed by inverse powers of N. As $N \to \infty$, only the first two terms in powers of t contribute, and produce an exponential function. This exponential function on the third line is quadratic in t, and is therefore the form of a Gaussian function. From years and years of Fourier and Laplace transforming, we know that the Fourier/Laplace transform of a Gaussian is a Gaussian (and inversely), and so we can immediately write down the probability distribution of $\tilde{X}$ in the large-N limit:

$$p(\tilde{X}) = \frac{1}{\sqrt{2\pi \sigma_x^2}} \, e^{-\frac{\tilde{X}^2}{2\sigma_x^2}} \,. \tag{2.46}$$

19
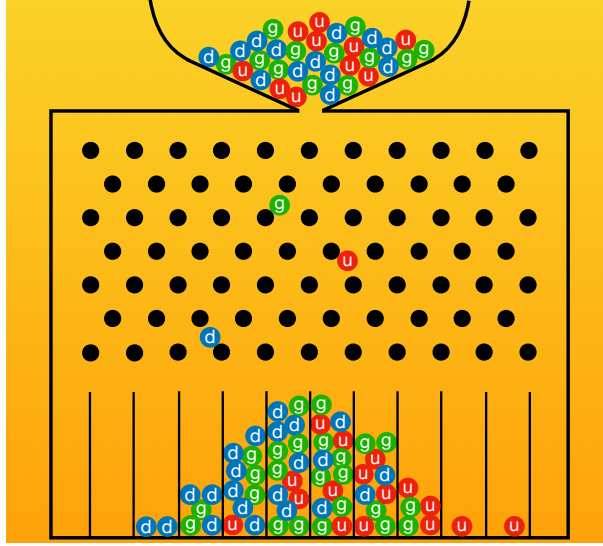
Figure 6: Illustration of a Galton board in which balls are dropped one-by-one from the hopper at top, and then strike the pegs as they fall, which randomly bump balls one space to the left or right.

Changing variables from $\tilde{X}$ to the original X (and remembering the Jacobian), we have

$$p(X) = \frac{1}{\sqrt{2\pi N \sigma_x^2}} e^{-\frac{(X - N\mu)^2}{2N\sigma_x^2}}, \tag{2.47}$$

as $N \to \infty$. This proves the central limit theorem, that the distribution of the sum of $N \to \infty$ i.i.d. random variables is a Gaussian with mean $\langle X \rangle = N\mu$ and variance $\sigma_X^2 = N\sigma_x^2$.

Fig. 6 illustrates a Galton board [115] (used in games like pachinko or plinko) which is perhaps the most direct way to observe the central limit theorem in action. Balls are dropped from the hopper at top one-by-one into the pegboard, and as they fall, they strike pegs which displaces the balls randomly one space left or right. Then, as the balls continue their fall, they accumulate some displacement from their original horizontal position from the peg strikes. For the Galton board of Fig. 6, the total horizontal displacement X in terms of peg spacing is

$$X = \sum_{i=1}^{7} x_i, \tag{2.48}$$

where $x_i \in \{-1, 1\}$ and there are 7 layers of pegs. Even though 7 is not an especially large number, the resulting distribution of balls in the bins at the bottom closely follows the Gaussian distribution. Actually, this Galton board represents the simplest version of a **random walk**, where at every step you randomly choose to move forward or backward. While on average, the displacement from the origin is 0, the standard deviation scales like $\sqrt{N}$ for N total steps, and so in any given instance of a random walk, you will typically find yourself $\sqrt{N}$ steps from the origin. In this particular figure of a Galton board, also note balls of different color and with letter labels; we will exploit this simple picture as a model for particle generation in Sec. 7.

2.3.2 An Infinitely Wide Neural Network as a Free Field Theory

There is a rather interesting connection between neural networks and statistical field theory, which here we only sketch. It has long been known that an infinitely wide neural network is a free field theory [116, 117, 118]. For some intuition, here we will consider just a single-layer network, but with an infinite number of neurons. Further, we will initialize the weights of the neurons from a Gaussian distribution with 0 mean and variance σ^2. From this starting point, we would like to determine the probability distribution of the response of this network, which we will denote as $p(\vec{x})$, where $\vec{x}$ is a formally infinite dimensional vector with entries corresponding to the individual neurons. Without loss of generality, we can write this probability as the exponential of some other function,

$$p(\vec{x}) \propto e^{-S(\vec{x})} . \tag{2.49}$$

In statistics, $S(\vec{x})$ is the negative logarithm of the probability, but in physics we would call it the action.

The "action" can be directly constructed from the assumption of Gaussian initialization, where

$$S(\vec{x}) = \frac{\vec{x} \cdot \vec{x}}{2\sigma^2} , \tag{2.50}$$

and so the probability distribution is

$$p(\vec{x}) = \frac{e^{-\frac{\vec{x} \cdot \vec{x}}{2\sigma^2}}}{Z} . \tag{2.51}$$

Here, Z is the appropriate normalization factor, or, in physics language, the partition function,

$$Z \equiv \lim_{N \to \infty} \int \prod_{i=1}^{N} \left[\frac{dx_i}{\sqrt{2\pi\sigma^2}} \right] e^{-\frac{\vec{x} \cdot \vec{x}}{2\sigma^2}} . \tag{2.52}$$

Just like in statistical physics, this partition function encodes everything you could ever want to ask about the distribution of neurons, $p(\vec{x})$. For example, to evaluate correlation functions between neurons, we introduce a vector source $\vec{J}$ into the partition function,

$$Z\left[\vec{J}\right] = \lim_{N \to \infty} \int \prod_{i=1}^{N} \left[\frac{dx_i}{\sqrt{2\pi\sigma^2}} \right] e^{-\frac{\vec{x} \cdot \vec{x}}{2\sigma^2} + \vec{J} \cdot \vec{x}} = e^{\frac{\sigma^2}{2} \vec{J} \cdot \vec{J}} . \tag{2.53}$$

The two-point connected correlator, for instance, is the second derivative with respect to $\vec{J}$ and then setting $\vec{J} = 0$:

$$\langle x_i x_j \rangle = \frac{\delta}{\delta J_i} \frac{\delta}{\delta J_j} Z\left[\vec{J}\right] \bigg|_{\vec{J}=0} = \delta_{ij} \, \sigma^2 . \tag{2.54}$$

This interpretation of an infinitely-wide network as a free, Gaussian field theory additionally illustrates that such networks do no learning. We can exactly evaluate the partition function, and can correspondingly evaluate any moment of the neuron response distribution we want, simply from initialization conditions. To go beyond the free-field limit, we must consider finite-width networks and additionally incorporate multiple layers. At finite width, effective four-point and

higher interactions are generated, but they can be treated perturbatively, through an expansion about the infinite width limit. Of course, interacting field theories in general do not have closed-form partition functions (at least for "realistic" theories that are non-integrable) and so honest learning or a systematic improvement in the neuron response must be considered in such cases. The statistics of finite-width networks is fascinating in its own right, and I cannot attempt to do it more justice here, but point the interested reader to Ref. [56] for explicit details.

3   Some Fundamental Ideas from Quantum Field Theory

With our machine learning results established, we now move on to reviewing some fundamental results in quantum field theory and QCD, which will correspondingly provide the foundation for actually calculating likelihood ratios and probability distributions on signal and background event classes that we can define. In the four binary discrimination examples that we study in the following sections, we will be guided by perturbation theory and so will need to define what to calculate and if it can be calculated at all. Additionally, we will work in the soft and/or collinear limit, narrowly focusing on small angle regions of our detector and where the description of particle dynamics significantly simplifies and factorizes from the rest of the event. Practically working in this collinear limit requires a jet algorithm to robustly define the region of interest, so we will review jet algorithms a bit here. However, in the following sections, all of our results will hold for any reasonable jet algorithm definition, so we will not need to specify the algorithm in particular.

3.1   Master Formula for Distributions from QFT

Our goal for these lectures is to determine the likelihood ratio for binary discrimination problems. As such, we need to calculate the probability distributions on signal and background events on which we have measured some collection of observables. Therefore, we need a master formula for calculating probability distributions in QCD and quantum field theory more generally, which of course was worked out nearly a century ago by Dirac and is referred to as **Fermi's Golden Rule** [119, 120]. We will actually need to slightly modify Fermi's Golden Rule from its usual form, because we want not only the probability of *something* happening, but we also want the probability that something happened with a particular measured value for an observable. In particular, Fermi's Golden Rule is typically stated as the rule for calculating scattering cross sections, while here we will need **differential cross sections**.

Fermi's Golden Rule for the differential cross section of some observable $\mathcal{O}$ on a class of events is

$$\frac{d\sigma}{d\mathcal{O}} = \int d\Phi \, |\mathcal{M}|^2 \, \delta\left(\mathcal{O} - \hat{\mathcal{O}}(\Phi)\right) . \tag{3.1}$$

This will be our master formula for making predictions in the sections that follow, so we want to ensure that we understand it deeply, in our bones. There are three parts of the calculation to consider. First, $d\Phi$ is **differential phase space** and represents the volume measure of relativistic, on-shell particle four-momentum subject to global energy and/or momentum constraints. Second, $|\mathcal{M}|^2$ is the squared **S-matrix element** that represents the probability density on phase space for the initial, prepared, state to transform into the final, observed, state. This is calculated with Feynman diagrams in perturbation theory. Third, $\delta\left(\mathcal{O} - \hat{\mathcal{O}}(\Phi)\right)$ is the measurement constraint,

where $\mathcal{O}$ is the measured value and $\hat{\mathcal{O}}(\Phi)$ is the functional form of the observable in phase space coordinates. Finally, the differential cross section has dimensions of a cross section (an area) divided by the dimensions of the observable $\mathcal{O}$. Without loss of generality, we can always work with dimensionless observables (which we do in these notes), and so to actually calculate the probability distribution of $\mathcal{O}$, we must normalize by the total cross section:

$$p(\mathcal{O}) = \frac{1}{\sigma} \frac{d\sigma}{d\mathcal{O}}, \tag{3.2}$$

where

$$\sigma = \int d\Phi \, |\mathcal{M}|^2 . \tag{3.3}$$

We will discuss each of the three main parts of the differential cross section calculation.

3.1.1 Differential Phase Space

We prepare the initial state, like scattering protons at a fixed energy at the LHC, but the final state observed in our detectors is completely determined by the whims and whys of quantum mechanics. As such, the particles that are produced must be allowed to have any energy or momentum consistent with the conservation laws of Nature. Further, the S-matrix assumes that particles are measured on the **celestial sphere**, points on the detector experiment that represent the direction of the particles formally an infinite distance and time away from the collision. We can thus build up the phase space for a total of N final state particles as follows. We assume we can only measure particles' momentum four-vectors and so differential phase space is differential in all particles' momenta:

$$d\Phi \supset \prod_{i=1}^{N} \left[d^4 p_i \right] . \tag{3.4}$$

Next, every detected particle must be on-shell (detection is classical) so the square of each four vector must be the appropriate particle mass:

$$d\Phi \supset \prod_{i=1}^{N} \left[d^4 p_i \, \delta(p_i^2 - m_i^2) \right] . \tag{3.5}$$

Next, energies (the p_0 component) must be non-negative, which we can enforce with a step function:

$$d\Phi \supset \prod_{i=1}^{N} \left[d^4 p_i \, \delta(p_i^2 - m_i^2) \, \Theta(p_{i,0}) \right] . \tag{3.6}$$

However, in the formulae that follow, we will not typically write the positive energy constraint explicitly, but just remember to include it in the bounds of integration. Next, we sprinkle factors of 2π around to account for Fourier transform normalization from position space to momentum space:

$$d\Phi \supset \prod_{i=1}^{N} \left[\frac{d^4 p_i}{(2\pi)^4} \, 2\pi \, \delta(p_i^2 - m_i^2) \right] . \tag{3.7}$$

As written so far, particle momenta are still allowed to be anything, and are unconstrained by any conservation laws, so we need to fix this up. However, once we actually write down the conservation law constraints, we have fixed ourselves to a particular frame. Phase space is necessarily Lorentz-invariant (because the universe is), but the concrete form that it takes requires energies and momenta evaluated in a particular frame. Here, we will just focus on two frames, the center-of-mass frame and, what we will study in great detail later, the **infinite momentum frame** [121, 122, 123, 124, 125, 126].

In the center-of-mass frame, the total energy is fixed and the total three-momentum is 0. There are thus four additional constraints to impose:

$$d\Phi_{\text{CM}} = \prod_{i=1}^{N} \left[\frac{d^4 p_i}{(2\pi)^4} \, 2\pi \, \delta(p_i^2 - m_i^2) \right] (2\pi)^4 \, \delta\left(Q - \sum_{i=1}^{N} p_{i,0}\right) \delta^{(3)}\left(\sum_{i=1}^{N} \vec{p}_i\right) . \qquad (3.8)$$

Here, the total energy of the event is Q. It is illustrative to determine the dimensionality of this phase space. First, there are $4N$ components of all N particle four-vectors, but they are all on-shell, so there are only $3N$ independent components. Energy-momentum conservation is an additional 4 constraints, and so the total dimensionality of N-body phase space is $3N - 4$. This result then gives you some appreciation for how challenging describing particle collision events is (at least naively). For events in which dozens of particles are produced, phase space can easily be 100 dimensional! Further, because of the non-linear relationship between energy and momentum, N-body phase space has a non-trivial topology, isomorphic to the product space of an $N - 1$ dimensional simplex and a $2N - 3$ dimensional sphere, $\Phi \simeq \Delta_{N-1} \times S^{2N-3}$ [127, 128, 129, 130, 131].[3] It has been demonstrated that machine learning on topologically non-trivial manifolds can have obstructions, unless information about the topology is included [137, 138, 139, 140, 141], so this is a feature of studying particle collision events that may be important to consider.

Infinite momentum or collinear phase space is rather different. Formally, we will expand phase space to lowest order in an expansion in relative angles between particles, thus considering the particles on phase space to effectively be collinear with one another. Because we work only to lowest order in the expansion, there is no Lorentz boost that can be performed to go back to the center-of-mass frame, for example. Mathematically, this means that we imagine particles to live exclusively on a tangent plane of the celestial sphere about a point fixed by the direction of the jet. This collinear limit is also called the infinite momentum frame because if we boost an event by infinite momentum in some direction (i.e., the boost velocity is the speed of light), then all of the particles will be exactly collinear with one another. Working with jets, which themselves are high-energy collimated sprays of particles, this phase space is the natural first approximation to work on, and will be the focus of the rest of these lectures.

To derive the collinear limit of phase space, we start with the full expression for N-body, massless phase space in the frame in which the particles have a total energy E with net momentum

---

[3]This explicit form for differential phase space is not the most convenient for actually doing calculations, however. A recursive formula that relates N to $N-1$ body phase space through integration over the possible energy that particle N can carry was first derived by Srivastava and Sudarshan [132]. George Sudarshan made enormous contributions to theoretical physics, but have come to be largely ignored, such as with the $V - A$ theory [133] (commonly credited to Feynman and Gell-Mann [134]) and the formulation of quantum optical phase space [135] (for which Roy Glauber won the 2005 Nobel Prize [136]).

along the $+\hat{z}$ axis:

$$d\Phi = \prod_{i=1}^{N} \left[\frac{d^4 p_i}{(2\pi)^4} \, 2\pi \, \delta(p_i^2) \right] (2\pi)^4 \, \delta\left(E - \sum_{i=1}^{N} p_{i,0} \right) \delta\left(p_z - \sum_{i=1}^{N} p_{i,z} \right) \delta^{(2)}\left(\sum_{i=1}^{N} \vec{p}_{i,\perp} \right). \tag{3.9}$$

In this expression, we have Lorentz-boosted the expression from the center-of-mass frame and so the total squared invariant mass is still Q^2:

$$Q^2 = E^2 - p_z^2. \tag{3.10}$$

To proceed, we will change variables to convenient coordinates in which the collinear limit is directly manifest. We define new momentum coordinates, called **light-cone coordinates**, that are defined with respect to the boost axis. We introduce the minus $-$, plus $+$, and perpendicular $\perp$ momenta as follows:

$$p^- \equiv p_0 + p_3\,, \qquad p^+ \equiv p_0 - p_3\,, \qquad \vec{p}_\perp \equiv (p_1, p_2)\,. \tag{3.11}$$

In light-cone coordinates, on-shell phase space is

$$d\Phi = \prod_{i=1}^{N} \left[\frac{dp_i^- \, dp_i^+ \, d^2 p_{i,\perp}}{2(2\pi)^3} \, \delta(p_i^- p_i^+ - p_{i,\perp}^2) \right] \tag{3.12}$$

$$\times (2\pi)^4 \, \delta\left(E - \sum_{i=1}^{N} \frac{p_i^- + p_i^+}{2} \right) \delta\left(\sqrt{E^2 - Q^2} - \sum_{i=1}^{N} \frac{p_i^- - p_i^+}{2} \right) \delta^{(2)}\left(\sum_{i=1}^{N} \vec{p}_{i,\perp} \right).$$

where the new factor of 2 is the Jacobian. We have also replaced the net momentum p_z by its relationship to the total energy and invariant mass.

Next, we now work in the collinear or infinite momentum space limit and here we will see the convenience of the light-cone coordinates. In terms of the angle θ of a particle from the boost axis (the $+\hat{z}$ direction), the light-cone momenta components are

$$p^- = p_0(1 + \cos\theta)\,, \qquad p^+ = p_0(1 - \cos\theta)\,, \qquad \vec{p}_\perp = p_0 \sin\theta (\cos\phi, \sin\phi)\,, \tag{3.13}$$

where ϕ is the azimuthal angle of the particle about the boost axis. In the collinear limit, $\theta \to 0$, these variables reduce to

$$p^- \to 2p_0\,, \qquad p^+ \to p_0 \frac{\theta^2}{2}\,, \qquad \vec{p}_\perp \to p_0\, \theta (\cos\phi, \sin\phi)\,. \tag{3.14}$$

It is useful to track the scaling with angle from the boost axis through the parameter $\lambda \sim \theta \ll 1$, and then we can expand phase space to leading power in λ everywhere to derive collinear or infinite momentum frame phase space. Note that with this scaling, the light-cone components scale like

$$p^- \sim \lambda^0\,, \qquad p^+ \sim \lambda^2\,, \qquad \vec{p}_\perp \sim \lambda\,. \tag{3.15}$$

This power counting and expansion is generally very useful for expansion and evaluation of Feynman integrals in gauge theories and is referred to as the **method of regions** [142, 143, 144, 145], by which one isolates different regions of an integral by asserting a power counting and uniformly expanding to leading power. Power counting is also the first step of defining effective field theories,

and this power counting in the collinear limit identifies and isolates the collinear modes of **soft-collinear effective theory** [146, 147, 148, 149] (for reviews, see Refs. [150, 151]), or SCET. We will apply power counting to deriving optimal discrimination observables from simple assumptions in Sec. 5.3.

The first thing we will do is to simplify the momentum-conservation δ-functions, expanding their arguments to leading power in the $\lambda \ll 1$ collinear limit. Note that p^+ is parametrically smaller than p^-, and so the energy conservation δ-function simplifies to

$$\delta\left(E - \sum_{i=1}^{N} \frac{p_i^- + p_i^+}{2}\right) \to \delta\left(E - \sum_{i=1}^{N} \frac{p_i^-}{2}\right). \tag{3.16}$$

Using this constraint, the second δ-function, conservation of p_z, simplifies to

$$\delta\left(\sqrt{E^2 - Q^2} - \sum_{i=1}^{N} \frac{p_i^- - p_i^+}{2}\right) \to \delta\left(\sqrt{E^2 - Q^2} - E + \sum_{i=1}^{N} \frac{p_i^+}{2}\right) \tag{3.17}$$

In the strict collinear limit, where all particles are truly collinear, their total invariant mass Q^2 would be 0, regardless of the energy E. Then, in the near-collinear limit, we take $Q \ll E$ and can expand the square-root factor as

$$\delta\left(\sqrt{E^2 - Q^2} - E + \sum_{i=1}^{N} \frac{p_i^+}{2}\right) \to \delta\left(\frac{Q^2}{2E} - \sum_{i=1}^{N} \frac{p_i^+}{2}\right). \tag{3.18}$$

Finally, we note that the transverse momentum δ-function

$$\delta^{(2)}\left(\sum_{i=1}^{N} \vec{p}_{i,\perp}\right) \tag{3.19}$$

is homogeneous in the power counting and no expansion can be done.

At this point, differential collinear phase space can be expressed as

$$d\Phi_{\text{coll}} \to \prod_{i=1}^{N} \left[\frac{dp_i^- \, dp_i^+ \, d^2 p_{i,\perp}}{2(2\pi)^3} \, \delta(p_i^- p_i^+ - p_{i,\perp}^2)\right] \tag{3.20}$$

$$\times (2\pi)^4 \, \delta\left(E - \sum_{i=1}^{N} \frac{p_i^-}{2}\right) \delta\left(\frac{Q^2}{2E} - \sum_{i=1}^{N} \frac{p_i^+}{2}\right) \delta^{(2)}\left(\sum_{i=1}^{N} \vec{p}_{i,\perp}\right).$$

We formally take the limit in which $\lambda \to 0$, and this correspondingly implies that $Q^2 \sim \lambda^2 \ll 1$. Naively, from the original expression for total $p_z = \sqrt{E^2 - Q^2}$, this would imply that $Q < E$, for example. However, taking our power counting and expansion seriously, to leading power, Q^2 loses all memory of relationship with total energy E, and is allowed to range over *all* values, $Q^2 \in [0, \infty)$. Further, once this expansion is taken, with total energy effectively infinite compared to the invariant mass, the invariant mass no longer constrains the particles, and should be integrated over. That is, to leading power in the collinear limit, there are no explicit angular scales. Then, integrating over the invariant mass Q^2, collinear phase space becomes

$$d\Phi_{\text{coll}} \to \int \frac{dQ^2}{2\pi} \prod_{i=1}^{N} \left[\frac{dp_i^- \, dp_i^+ \, d^2 p_{i,\perp}}{2(2\pi)^3} \, \delta(p_i^- p_i^+ - p_{i,\perp}^2)\right] \tag{3.21}$$

$$\times (2\pi)^4 \, \delta \left(E - \sum_{i=1}^{N} \frac{p_i^-}{2} \right) \delta \left(\frac{Q^2}{2E} - \sum_{i=1}^{N} \frac{p_i^+}{2} \right) \delta^{(2)} \left(\sum_{i=1}^{N} \vec{p}_{i,\perp} \right)$$

$$= 2(2\pi)^3 E \prod_{i=1}^{N} \left[\frac{dp_i^- \, dp_i^+ \, d^2 p_{i,\perp}}{2(2\pi)^3} \delta(p_i^- p_i^+ - p_{i,\perp}^2) \right] \delta \left(E - \sum_{i=1}^{N} \frac{p_i^-}{2} \right) \delta^{(2)} \left(\sum_{i=1}^{N} \vec{p}_{i,\perp} \right) .$$

We do the integral over the total squared invariant mass and divide by 2π because we effectively have to "undo" the on-shell δ-function that would impose $Q^2 = E^2 - p_z^2$.

We now just have a couple more things to tidy up. First, as all p^+ dependence in the momentum conservation δ-functions has been removed, let's integrate over all p^+ components with the on-shell δ-functions. This produces

$$d\Phi_{\text{coll}} \to 2(2\pi)^3 E \prod_{i=1}^{N} \left[\frac{dp_i^- \, d^2 p_{i,\perp}}{2 p_i^- (2\pi)^3} \right] \delta \left(E - \sum_{i=1}^{N} \frac{p_i^-}{2} \right) \delta^{(2)} \left(\sum_{i=1}^{N} \vec{p}_{i,\perp} \right) . \tag{3.22}$$

We can also integrate over particle N's transverse momentum with the appropriate δ-function:

$$d\Phi_{\text{coll}} \to E \prod_{i=1}^{N-1} \left[\frac{dp_i^- \, d^2 p_{i,\perp}}{2 p_i^- (2\pi)^3} \right] \frac{dp_N^-}{p_N^-} \delta \left(E - \sum_{i=1}^{N} \frac{p_i^-}{2} \right) . \tag{3.23}$$

We can also introduce some convenient coordinates for the final, compact expression. First, we will introduce the energy fractions

$$z_i \equiv \frac{p_i^-}{2E} , \tag{3.24}$$

and correspondingly express the transverse momentum integration measure in polar coordinates:

$$d^2 p_{i,\perp} = p_{i,\perp} \, dp_{i,\perp} \, d\phi_i = z_i^2 E^2 \, \theta_i \, d\theta_i \, d\phi_i . \tag{3.25}$$

With these expressions, collinear phase space takes the nice form, where

$$d\Phi_{\text{coll}} = \frac{E^{2(N-1)}}{2^{N-1} (2\pi)^{3N-3}} \prod_{i=1}^{N-1} [dz_i \, z_i \, \theta_i \, d\theta_i \, d\phi_i] \frac{dz_N}{z_N} \delta \left(1 - \sum_{i=1}^{N} z_i \right) . \tag{3.26}$$

Recall that the polar angle θ_i is the angle of particle i from the boost axis, the $+\hat{z}$ direction, and ϕ_i is its azimuthal angle about the boost axis.

3.1.2 Two-Body Collinear Phase Space

As an explicit example that we will revisit over and over in these lectures, let's evaluate this phase space for jets with two particles. With $N = 2$, phase space simplifies to

$$d\Phi_{\text{coll}}^{(N=2)} = \frac{E^2}{2(2\pi)^3} \, dz_1 \, z_1 \, \theta_1 \, d\theta_1 \, d\phi_1 \, \frac{dz_2}{z_2} \, \delta \left(1 - z_1 - z_2 \right) . \tag{3.27}$$

We have drawn a representation of a jet with two particles in Fig. 7, both as vectors and as points on a tangent plane of the celestial sphere. With some additional assumptions, this phase space can
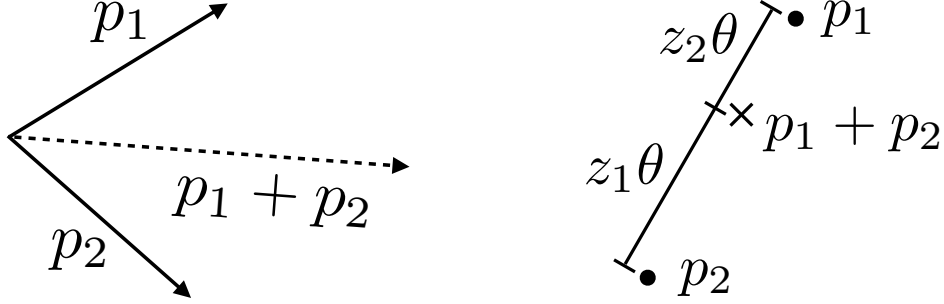
Figure 7: An illustration of a jet with two collinear particles. Left: Representation of momentum vectors of particles 1 and 2 and their sum. Right: Representation of the direction of momentum of particles 1 and 2 as points on the two-dimensional celestial sphere, and the direction of their sum.

be nicely expressed in a very compact form. If we assume that there is an azimuthal symmetry about the boost axis, then we can integrate over the azimuthal angle and produce

$$d\Phi_{\text{coll}}^{(N=2)} = \frac{E^2}{8\pi^2} \, dz_1 \, z_1 \, \theta_1 \, d\theta_1 \, \frac{dz_2}{z_2} \, \delta \left(1 - z_1 - z_2\right) . \tag{3.28}$$

Next, we can express the angle θ_1 of particle 1 with respect to the boost axis instead as the angle θ that separates particles 1 and 2. By conservation of transverse momentum in the collinear limit (and illustrated in Fig. 7), this angle is

$$\theta_1 = z_2 \theta . \tag{3.29}$$

Then, this two-body collinear phase space is

$$d\Phi_{\text{coll}}^{(N=2)} = \frac{E^2}{8\pi^2} \, z(1 - z) \, dz \, \theta \, d\theta . \tag{3.30}$$

Here z is an energy fraction of one of the final state collinear particles, $z \in [0, 1]$.

This is often re-expressed in terms of the energy fraction z and the invariant mass of the two particles, s. In the collinear limit, this invariant mass is

$$s = z(1 - z)E^2\theta^2 , \tag{3.31}$$

and the corresponding expression for collinear phase space is

$$d\Phi_{\text{coll}}^{(N=2)} = \frac{dz \, ds}{(4\pi)^2} , \tag{3.32}$$

which is exceptionally compact. We will study this phase space from a rather different perspective and set of initial assumptions in Sec. 4.

3.2 Collinear and Soft Factorization

QCD is a gauge theory for which there is a spin-1, massless boson, the gluon, that is ultimately responsible for the forces that color charged particles act on one another. In this way, QCD is analogous to electromagnetism for which the photon is exchanged between electrically charged particles

and correspondingly leads to Coulomb's law, the Biot-Savart law, and the full weight of Maxwell's equations. Phenomenologically, we also know that electromagnetism has many interesting limits, such as acting coherently like a wave at long distances, or, by Fourier transformation, at low energies, or when a charged particle is rapidly accelerated, it emits most of its power in radiation that is approximately collinear to the particle's direction of motion. These phenomena are also present in QCD, and, especially the latter, is what is ultimately responsible for jets.

Further, it isn't so much that QCD has an analogous Larmor formula (which it does, by the way), but rather that these soft and collinear limits **factorize** from the dynamics of whatever else is going on in the event of interest. This factorization or separation of a squared matrix element into a product of factors enables a self-similarity to the emission structure of QCD events. Radiation collinear to a high-energy quark, for instance, factorizes from the dynamics of the quark, which can then emit more collinear radiation that factorizes from the quark, etc., producing a whole shower or cascade of collinear particles, manifesting itself as a jet. We will actually start from this scale-invariant assumption in the following section and see how far its consequences go, but here we must first review this soft and collinear factorization and present the explicit forms of the squared matrix elements that we will manipulate and study in these notes.

3.2.1   Collinear Splitting Factorization

In a theory of massless particles, such as QCD at high energies where quark masses can be ignored, propagators in the calculation of Feynman diagrams can diverge in one of two ways. This can be directly observed by simply writing down the expression for a propagator formed from the sum of two massless momenta p_1, p_2:

$$\frac{1}{(p_1 + p_2)^2} = \frac{1}{2 p_1 \cdot p_2} = \frac{1}{2 E_1 E_2 (1 - \cos \theta_{12})} \, , \tag{3.33}$$

where E_1, E_2 are the energies of the two momenta and θ_{12} is their relative angle. As the focus of this subsection, two massless particles, for example, can become collinear in which they travel in the same direction. In this case, their invariant mass (the inverse of the propagator) vanishes because two massless particles traveling in the same direction act as an effective meta-massless particle itself. The fact that the propagator diverges in the collinear limit means that such a configuration is likely to occur; particles *want* to be collinear with one another. We will address the soft limit of the propagator, when, say, $E_2 \to 0$, in the next subsection.

Indeed, these divergences can be interpreted through another lens. Feynman diagram perturbation theory is a degenerate perturbation theory and as a degenerate perturbation theory, we know that divergences arise in intermediate calculations, but cancel when all exactly degenerate states are inclusively summed over. We will revisit this point again and again, but for now, this interpretation means that collinear divergences in Feynman diagram perturbation theory correspond to states that are indistinguishable from other states in the theory with a different number of collinear particles. That is, the quantum numbers of exactly collinear particles in QCD correspond to the quantum numbers of individual quarks and gluons, and so by the rules of degenerate perturbation theory, we must sum over states with any possible number of collinear particles. That is, there is no experiment that you can perform to determine the number of exactly collinear particles produced in any process, because all you can measure are net quantum numbers. We will see how to do this in our first example next lecture, but the starting point for this analysis is to determine the squared matrix element in the collinear limit.
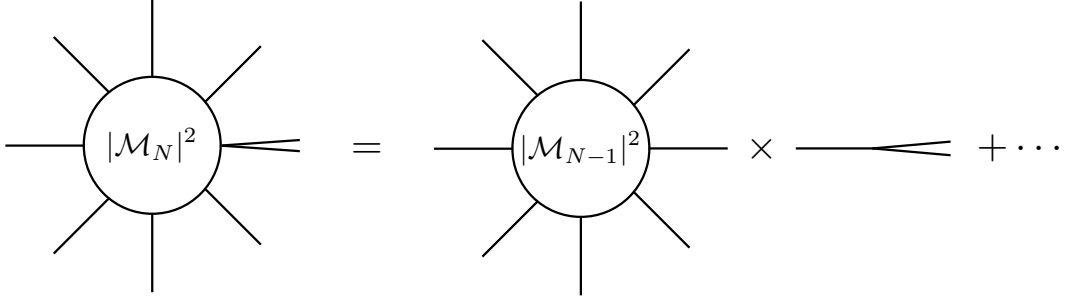
Figure 8: Schematic illustration of the collinear limit of an initial N point square S-matrix element $|\mathcal{M}_N|^2$ that factorizes into a product of an $N-1$ point squared matrix element $|\mathcal{M}_{N-1}|^2$ and the collinear splitting function (denoted as the $1 \to 2$ splitting at right).

From the early days of QCD, and from analogy with electromagnetism, it has been known that squared matrix elements factorize in the collinear limit. A squared matrix element of N final state particles $|\mathcal{M}(p_1, \ldots, p_N)|^2$ calculated in perturbation theory takes the following form when two particles i and j become collinear:

$$|\mathcal{M}(p_1, \ldots, p_i, p_j, \ldots, p_N)|^2 \big|_{i \| j} = \frac{8\pi\alpha_s}{s_{ij}} P_{(ij) \to ij}(z) \, |\mathcal{M}(p_1, \ldots, p_{(ij)}, \ldots, p_N)|^2 + \cdots . \qquad (3.34)$$

By "collinear", I mean that the angle between particles i and j is much, much smaller than any other angle between particles i and j and other particles in the event. At right, $|\mathcal{M}(p_1, \ldots, p_{(ij)}, \ldots, p_N)|^2$ is the squared matrix element in which momenta of individual particles i and j are replaced by their sum $p_{(ij)} = p_i + p_j$. The factor multiplying this squared matrix element has the propagator $1/s_{ij}$ explicit, as well as the coupling α_s that controls the probability to split at all. Ellipses hide terms that are less singular in the collinear limit (i.e., do not have the factor of the propagator). This equation is drawn schematically in Fig. 8.

The most interesting part of this formula, however, are the **splitting functions**, $P_{(ij) \to ij}$ which describe the probability of the distribution of energy fraction z carried by particle i in the splitting. These are **universal** in that they are independent of the squared matrix element or process that one considers, and are intrinsic to the theory of QCD. All possible $1 \to 2$ splitting functions are referred to as the **Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) kernels** [152, 153, 154, 155, 156]:

$$P_{q \to qg}(z) = C_F \frac{1 + z^2}{1 - z} \,, \qquad\qquad P_{q \to gq}(z) = C_F \frac{1 + (1 - z)^2}{z} \,, \qquad (3.35)$$

$$P_{g \to q\bar{q}}(z) = T_R \left[z^2 + (1 - z)^2 \right] \,, \qquad P_{g \to gg}(z) = C_A \left(\frac{z}{1 - z} + \frac{1 - z}{z} + z(1 - z) \right) .$$

These splitting functions are further controlled by color factors C_F, C_A, and T_R which quantify how the three colors of QCD are shared between the quarks and gluons involved in the splitting. C_F and C_A are the quadratic Casimir operators of the fundamental and adjoint representations of SU(3) color, respectively, and T_R is the normalization of the fundamental representation or Gell-Mann matrices of QCD (i.e., the Killing form). In an SU(N) gauge theory, the quadratic Casimirs are

$$C_F = \frac{N^2 - 1}{2N} \,, \qquad\qquad C_A = N \,, \qquad (3.36)$$

and so are $C_F = 4/3$ and $C_A = 3$ in QCD. By convention, we also fix $T_R = 1/2$.

Note, however, that these splitting functions are in general not integrable on $z \in [0, 1]$, but typically diverge if $z \to 0, 1$. The additional soft divergences correspond to a gluon produced in the splitting carrying 0 energy, which again is a further degenerate state. (More on this in a second.) Individual quarks, on the other hand, have no corresponding soft divergence, for a few reasons. For example, quarks are fermions and carry spin-1/2 and there is no possible degenerate state in which a spin-1/2 particle becomes soft and "disappears". By angular momentum conservation, you will always know if a quark has become soft.

We will almost exclusively only consider $1 \to 2$ collinear splittings in our analyses that follow, but in general, $1 \to n$ collinear splitting functions exist and can be calculated. For this more general splitting, the factorization in the collinear limit takes basically the same form as Eq. (3.34), with the factors of the propagator and powers of the coupling modified to describe additional particle emission. Concretely, the $1 \to 3$ splitting functions have been worked out [157, 158], and the factorization in this limit takes the form:

$$
|\mathcal{M}(p_1, \ldots, p_i, p_j, p_k, \ldots, p_N)|^2\big|_{i\|j\|k} \tag{3.37}
$$
$$
= \left(\frac{8\pi\alpha_s}{s_{ijk}}\right)^2 P_{(ijk)\to ijk}(\{z\}, \{s\}) \, |\mathcal{M}(p_1, \ldots, p_{(ijk)}, \ldots, p_N)|^2 + \cdots,
$$

for three collinear particles i, j, k. The novel aspect starting at $1 \to 3$ splitting is that the splitting function $P_{(ijk)\to ijk}(\{z\}, \{s\})$ depends on three energy fractions, z_i, z_j, z_k with $z_i + z_j + z_k = 1$, and, in general, all pairwise invariant masses s_{ij}, s_{ik}, s_{jk} with $s_{ij} + s_{ik} + s_{jk} = s_{ijk}$.[4]

3.2.2 Soft Gluon Factorization

The second possibility for particle configuration in which a propagator diverges is if a soft or low energy gluon is emitted at arbitrary angles to the other, harder, particles in the event. For jet physics specifically, in which we focus on the collinear, high energy dynamics, soft, wide-angle gluon emission is less directly relevant (at least for these lectures), but we include its discussion for completeness and we will encounter a few scenarios where we will need to understand why and where gluons become soft. As hinted at in introducing this section, the soft limit by Fourier transformation corresponds to the long distance or classical limit. Soft divergences have a bit of a different interpretation than collinear divergences, as well. There is no way to determine the number of arbitrarily soft gluons produced because they carry no energy and so never "light" up any physical detector. However, with long wavelength, they are sensitive to the global flow of color charge and act coherently and so conspire in the large number limit as a classical wave.

In the limit that a gluon k carries arbitrarily low energy, also called the **eikonal limit**, the matrix element factorizes as [169, 170, 171, 172]

$$
|\mathcal{M}(k, p_1, \ldots, p_N)|^2\big|_{k\to 0} = -8\pi\alpha_s \sum_{i,j=1}^{N} \mathbf{T}_i \cdot \mathbf{T}_j \frac{s_{ij}}{s_{ik} s_{kj}} |\mathcal{M}(p_1, \ldots, p_N)|^2 + \cdots. \tag{3.38}
$$

This has a few more moving parts than for a collinear splitting, because of the global sensitivity of soft gluons. First, the sum runs over all hard particles in the event i, j, including $i < j$, $i > j$, and

---

[4]For a complete collinear calculation at this order in perturbation theory, one also needs the one-loop, $1 \to 2$ splitting functions [159, 160, 161]. Even higher order splitting functions are now known, as well [162, 163, 164, 165, 166, 167, 168].

$i = j$. The propagator, or eikonal factor, is the ratio of invariant masses $s_{ij}/s_{ik}s_{kj}$, and accounts for the divergent propagator if the soft gluon k is emitted from either particle i or j.

This form of the factorization takes the form of a sum over **dipoles**, for exactly the same reason that an accelerating dipole in electromagnetism emits radiation. Sensitivity to electric charge in a radiation formula is replaced by sensitivity to the product of color matrices $\mathbf{T}_i \cdot \mathbf{T}_j$ that describes how particles i and j share color; i.e., the color that flows along the dipole formed by particles i and j. This completely inclusive double sum is convenient to represent the sum over dipoles because it nicely manifests color conservation. We have that

$$\sum_{i,j=1}^{N} \mathbf{T}_i \cdot \mathbf{T}_j = \left(\sum_{i=1}^{N} \mathbf{T}_i \right)^2 = 0 \,, \tag{3.39}$$

but the total color charge of *any* event is 0 (summing over all initial and final state particles), because interactions cannot create or annihilate net charge. To actually evaluate the color matrix dot products can be tricky in general, and typically requires working in an explicit color basis and then simply keeping track of all color matrices in the evaluation of Feynman diagrams.[5] However, for low point amplitudes, with 2 or 3 quarks or gluons, the color factors can be evaluated exactly simply from color conservation and the square of color matrices. Specifically, the square of the color matrix $\mathbf{T}_i$ is the corresponding quadratic Casimir of particle i:

$$\mathbf{T}_i^2 = C_i \,. \tag{3.40}$$

It is correspondingly useful to validate consistency of the soft and collinear limit from this factorization, with the expressions from collinear factorization earlier. Let's say that soft gluon k is collinear with particle i. Then, note that the invariant mass of k with another particle $j \neq i$ can be expressed as

$$s_{kj} = z s_{ij} \,, \tag{3.41}$$

where $z \ll 1$ is the energy fraction of gluon k with respect to the energy of particle i. Then, the factorization of the squared matrix element becomes

$$|\mathcal{M}(k, p_1, \ldots, p_N)|^2 \big|_{k \to 0, k \| i} = -\frac{16\pi\alpha_s}{z s_{ik}} |\mathcal{M}(p_1, \ldots, p_N)|^2 \sum_{j=1, j\neq i}^{N} \mathbf{T}_i \cdot \mathbf{T}_j + \cdots \tag{3.42}$$

$$= -\frac{16\pi\alpha_s}{z s_{ik}} |\mathcal{M}(p_1, \ldots, p_N)|^2 \, \mathbf{T}_i \cdot \sum_{j=1, j\neq i}^{N} \mathbf{T}_j + \cdots$$

$$= \frac{16\pi\alpha_s}{z s_{ik}} \mathbf{T}_i^2 |\mathcal{M}(p_1, \ldots, p_N)|^2 + \cdots \,,$$

where we used color conservation on the third line. Note that we have multiplied by a factor of 2 because the original sum over dipole particles was completely inclusive in both i and j. This indeed agrees with the soft limit of the collinear factorization formulas we presented, with the ellipses hiding terms that are higher order in the soft and/or collinear limits.

---

[5]A very efficient organizing principle for color management is the topological expansion of 't Hooft [173, 174]. Color-ordering and diagrammatic computation of traces of products of color matrices is a standard practice for multi-loop or multi-leg amplitude calculations [175, 176, 177, 178, 179, 180].

In the past decade or so, there has been a flurry of interest in the next-to-leading soft limit behavior of scattering amplitudes in gauge theories, like QCD, and gravity, see, e.g., Refs. [181, 182, 183, 184, 185, 186, 187, 188]. In gauge theories, the next term in the soft gluon expansion is referred to as the Low-Burnett-Kroll theorem [169, 171], and is sensitive to the angular momentum of the two particles i and j in the dipole off of which the soft gluon is emitted. However, unlike the leading, eikonal limit, the Low-Burnett-Kroll term only holds for tree-level amplitudes and does not generalize simply for arbitrary loop amplitudes for the simple reason that real emission soft divergences start to mix with soft and collinear virtual divergences starting at one-loop order. Loop level subleading soft theorems can systematically be constructed [189, 190], but they are rather a mess and bookkeeping becomes a serious issue. At any rate, we will only need to understand the leading soft behavior here.

3.3 Infrared and Collinear Safety

In identifying the limits in which collinear and soft factorization of squared matrix elements exist, we uncovered a problem for interpretation of our master formula. Feynman diagram perturbation theory is degenerate, and this manifests as honest non-integrable divergences of squared matrix elements in the soft and/or collinear limit. Non-integrability kills hope of interpretation of our master formula as a probability distribution because, if anything, probabilities are integrable. However, we also know that this cannot be the whole story because degenerate perturbation theory is only problematic if you consider a subset of all possible degenerate states. Again, summing over all degenerate states eliminates the divergences and renders predictions finite and, correspondingly, integrable. So, for our master formula to be useful, we need to ensure that all degenerate states are indeed inclusively summed over.

The first step in doing this is to identify what these degenerate states are in our perturbation theory of QCD. In the previous section, we discussed *real* particle emission in the collinear or soft limit. These real contributions are positive because they correspond to honest absolute squared amplitudes. So, if their divergences are to cancel, then there must be a negative divergent contribution to account for. This negative contribution is from, of course, the virtual, or loop, diagrams. An N-point amplitude has a loop expansion of the form

$$\mathcal{M}(p_1, \ldots, p_N) = \mathcal{M}^{(0)}(p_1, \ldots, p_N) + \alpha_s \mathcal{M}^{(1)}(p_1, \ldots, p_N) + \cdots, \tag{3.43}$$

where the superscript (l) denotes the number of loops in the diagram; e.g., (0) is the tree-level diagram and (1) is the one-loop diagram. In our master formula we need the absolute square of this amplitude, which, through one-loop is

$$|\mathcal{M}(p_1, \ldots, p_N)|^2 = \left(\mathcal{M}^{(0)}(p_1, \ldots, p_N) + \alpha_s \mathcal{M}^{(1)}(p_1, \ldots, p_N) + \cdots \right) \tag{3.44}$$
$$\times \left(\mathcal{M}^{(0)}(p_1, \ldots, p_N) + \alpha_s \mathcal{M}^{(1)}(p_1, \ldots, p_N) + \cdots \right)^*$$
$$= |\mathcal{M}^{(0)}(p_1, \ldots, p_N)|^2 + 2\alpha_s \operatorname{Re}\left(\mathcal{M}^{(0)}(p_1, \ldots, p_N)\mathcal{M}^{(1)}(p_1, \ldots, p_N)^* \right) + \cdots.$$

Simply matching orders in α_s, the virtual contribution, the real part of the product of the tree-level and conjugate of the one-loop amplitudes, must be negative if there is any hope of divergence cancellation.

Additionally, this loop expansion produces no additional particles, and so all of the terms in this loop expansion are to be evaluated on N-body phase space. By contrast, a collinear splitting from

an N-point amplitude produces an additional real particle, and so are to be evaluated on $N+1$ body phase space. Actual infinities are only exactly located at points on phase space where particles are exactly collinear, where there is perfect confusion over how many particles are produced collinearly. Thus, for the real infinities on $N + 1$ body phase space from collinear or soft emission to cancel with the virtual infinities of N point loop amplitudes, we must ensure that any observable that we measure on these events cannot individually identify exactly collinear or exactly 0 energy particles.

This requirement for the success of degenerate Feynman diagram perturbation theory is called **infrared and collinear safety** or IRC safety, and has historical understanding and formalization in the Bloch-Nordseieck theorem of QED and the Kinoshita-Lee-Nauenberg (KLN) theorem more generally in any quantum field theory [191, 192, 193]. The first use of IRC safety in QCD was to provide a robust, and correspondingly calculable, definition of jets [194], which is now referred to as the Sterman-Weinberg algorithm.[6] We will review jet algorithms in the following section, but while the Sterman-Weinberg algorithm was extremely important historically, it is not very practical and is therefore not used experimentally, as it defines jets as containing a fraction $\epsilon > 0$ of the total energy within an angular size $\delta > 0$. Nevertheless, its IRC safety is easy to understand, because exactly collinear particles are always within any angle $\delta > 0$ and exactly 0 energy particles always have less than any fraction $\epsilon > 0$ of the total energy of an event. So, the Sterman-Weinberg jet definition inclusively sums over degenerate states.

A modern definition of IRC safety for a general observable $\hat{\mathcal{O}}(\Phi)$ is typically stated as [49]:

> The observable $\hat{\mathcal{O}}(\Phi)$ is IRC safe if, for any three-momentum of a particle $\vec{p}_i$ on phase space Φ, the observable's value is invariant under the splitting $\vec{p}_i \to \vec{p}_j + \vec{p}_k$ if the momenta $\vec{p}_j$ and $\vec{p}_k$ are parallel or one of them is small.

This will be our guiding principle in constructing observables for which our master formula can be applied, and for which we can make systematically-improvable predictions. In particular, a very simple observable structure that is immediately IRC safe is if the observable is only sensitive to linear sums of particle three-momenta. While I know no proof of this statement, in practice this linear sum rule seems to nearly be necessary and sufficient for IRC safety and even some of the rather complicated observables we will discuss later can all be expressed as a function of sums of momenta, even if we write them for compactness in a rather different way.[7] One can then say the art of jet substructure is in the construction of observables that are both sensitive to the phenomena of interest as well as IRC safe for predictability.

However, it must be said in the same breath that there is nothing "bad" or "undesirable" about observables that are not IRC safe. Perhaps the simplest such observable is the total particle multiplicity (just the number of particles produced), as an arbitrarily soft emitted particle will change this quantity by one unit. IRC unsafe observables are simply those for which our master formula does not apply, but we will nevertheless need an understanding of them. As these lectures progress, we will stray further and further from the light of IRC safe observables, and find strange situations in which our master formula secretly works for IRC unsafe observables, and for others, like multiplicity, where we simply have to throw up our hands and attack the problem from a

---

[6]See Refs. [195, 196, 197, 198, 199, 200] for earlier work on cancellation of infrared divergences in QCD and non-Abelian gauge theories.

[7]A related statement is proved, however. Multilinear symmetric functions of particle momenta span the space of IRC safe observables, see Refs. [201, 202, 203, 204, 205], in the sense that arbitrary linear combinations of them can approximate any IRC safe observable with arbitrary accuracy. The statement that I know of no proof for is if there exist IRC safe observables that are expressed as functions that are irreducibly not linear in particle momenta.

completely different direction. To emphasize again, our master formula and IRC safe observables have a use and are interesting, but the world is much larger so we will need more and powerful tools. In fact, we will see when IRC safety can guide us no further, in many cases the central limit theorem is there to lead on.

3.3.1 Master Formula for Jet Distributions

With phase space, factorization, and IRC safety understood, we can now write down a master Fermi's Golden Rule for jet substructure. The formula we will return to over and over as a first point of understanding in calculating distributions of observables $\mathcal{O}$ on jets with two particles in the collinear limit. For many of the observables we study, this is the first interesting case, or leading-order in the coupling when distributions are non-trivial. This master formula is then

$$
\begin{aligned}
\frac{d\sigma^{(0)}}{d\mathcal{O}} &= \int d\Phi_{\text{coll}}^{(N=2)} \frac{8\pi\alpha_s}{s} P_{(ij)\to ij}(z)\, \delta\left(\mathcal{O} - \hat{\mathcal{O}}(\Phi_{\text{coll}}^{(N=2)})\right) \\
&= \frac{\alpha_s}{\pi} \int \frac{d\theta}{\theta}\, dz\, P_{(ij)\to ij}(z)\, \delta\left(\mathcal{O} - \hat{\mathcal{O}}(z,\theta)\right).
\end{aligned}
\tag{3.45}
$$

3.4 Aside: Jet Algorithms

Given a proton collision event at the LHC, like that displayed in Fig. 1, we can see by eye the jets as the high-energy, collimated deposits in the experimental detector. However, at the extreme data collection rates of 1000 recorded events per second (or more), we can't simply have humans look at every event display and circle the jets for further analysis. Instead, we need a procedure or algorithm for identifying jets in an event that clusters particles that are close in angle and correlated in energy, and further, is ideally IRC safe so that we can calculate rates for jet production in perturbation theory. Due to the soft and collinear divergences in QCD, jets in experiment are highly correlated with a single initiating parton that subsequently radiated many particles that were ultimately detected. Actually, given a jet algorithm and unlabeled particle physics data, the clustering of particles in that data is an example of **unsupervised learning** in machine learning because the algorithm identifies structure without human intervention or explicit ground truth correlation. The jets that a jet algorithm finds are then proxies for individual quarks and gluons produced in the short distance collision event.

Historically, there have been many jet algorithms defined and used (we already mentioned the Sterman-Weinberg algorithm), but as theoretical and experimental goals sharpened, one class of jet algorithm has come to be standard. **Pairwise clustering** or **recombination algorithms** proceed by building structure in the event by clustering together pairs of particles that are closest together by a particular measure until some termination criteria is met. In this way, such algorithms effectively undo the collinear splitting process, and so, at least in the strict collinear limit, can be understood in this way. However, away from exact collinearity, there is significant freedom in how this clustering is actually done.

The specific algorithms for pairwise clustering were developed in the early 1990s but in some sense was "perfected" in 2008 [206, 207, 208, 209, 210, 211].[8] They consist of two, logically distinct, parts, the **clustering metric** and the **recombination scheme**. The clustering metric defines *which* particles are combined while the recombination scheme defines *how* particles are combined.

---

[8]See Refs. [212, 213, 214, 215, 216] for earlier examples of pairwise clustering algorithms.

We will first present the recombination algorithm, and through it discuss the different metrics and recombination schemes that are used.

3.4.1 Clustering Metric

The specific algorithm for pairwise reclustering is as follows:

1. Calculate the clustering metric

$$d_{ij}^{(n)} = \min \left[E_i^{2n}, E_j^{2n} \right] \frac{\theta_{ij}^2}{R^2} \tag{3.46}$$

   between all pair of particles i, j in the event. Here, n is a parameter that defines the particular algorithm. The most common choices are $n = 1$ which is the k_T **algorithm**, $n = 0$ which is the **Cambridge/Aachen (C/A) algorithm**, and $n = -1$ which is the **anti-k_T algorithm**. In general, this clustering metric defines a class of algorithms that are referred to as k_T-type.

2. For the pair of particles with the smallest $d_{ij}^{(n)}$, combine their momentum into a meta-particle (ij). The standard way to do this is through **E-scheme** recombination by which the four-vectors are summed:

$$p_{(ij)} = p_i + p_j \,. \tag{3.47}$$

   We will say more about other recombination schemes shortly. Then, remove i and j from the list of particles, and add (ij) in their place.

3. Continue this procedure, recalculating the pairwise metric $d_{ij}^{(n)}$ and clustering the smallest pair. In addition, also calculate the "beam" metric

$$d_{ib}^{(n)} = E_i^{2n} \,. \tag{3.48}$$

   If $d_{ib}^{(n)}$ is smaller than all of the $d_{ij}^{(n)}$, then remove all particles from the event list that had been clustered into i and call them jet i. Once all jets are clustered in this way, the algorithm terminates.

Specifically, the algorithm presented here uses energies and angles for simplicity of expressions, but these should be modified to transverse momenta and azimuth-(pseudo)rapidity distances at a hadron collider. This algorithm produces jets with a characteristic angular size of R; that is, all particles within an angle of about R of hard, collinear radiation are put into a single jet.

k_T-type pairwise clustering algorithms are IRC safe. Low energy particles don't affect the clustering sequence and further don't affect global properties of jets, like their total energy or direction of net momentum. Collinear particles are always clustered in the same jet because of the pairwise angular weighting factor in the clustering metric. So, as long as you ask IRC safe questions about jets or their substructure, k_T-type jets are themselves IRC safe. However, historically, it was not always the case that IRC safe algorithms were used in experiment. For a few decades, from the early QCD era in the 1970s to really until the late 20-aughts (more about why below), experiments

typically clustered jets with iterative, seeded cone algorithms (see, e.g., Refs. [217, 218, 219, 220]).[9] A seeded cone algorithm starts with a seed (some particle in the event), and places a cone, or disk of radius R, centered about the seed. Then, the net momentum of the particles in the jet is calculated, and the jet center moved to the location of net momentum. This procedure continues until the distance that the cone jet center moves is less than some criteria. Seeded cone algorithms are not IRC safe because arbitrarily soft particles can seed a cone and, by chance, cluster two hard particles that are further than R apart in angle.

One thing to note about jet algorithms in QCD in general is that there is no natural value for the jet radius R. QCD at high energies is an approximately scale-invariant theory, and correspondingly has no intrinsic angular scales. Therefore, regardless of the value of $R > 0$ you use to cluster jets, you will always find jets! In practice, however, experimental detectors break scale invariance, and experimentalists weigh several factors to determine an optimal jet radius, including particle resolution, average jet energy from fixed collision energy, maximal pseudorapidity range, etc. For all-purpose jet finding, the radius that is typically used at both ATLAS and CMS is about $R = 0.4$, but for some applications, a large or "fat" jet radius of $R \sim 1$ may be used. This latter case is especially useful when there is a known angular scale for the jets of interest; namely, those jets that indeed have an intrinsic mass. Such an example is in searching for energetic top quarks that decay hadronically, which have, because of the large mass of top quarks, $m_t \approx 173$ GeV, a characteristic angular size of about $R \sim 2m_t/E$.

The k_T and C/A algorithms tend to cluster soft, wide angle particles first, and then cluster those structures with hard, collinear radiation. As such, the shape of k_T or C/A jets is highly irregular, which is undesirable for experimental calibration of such jets (and one reason for significant foot-dragging by experiments to adopt such algorithms). However, in February 2008, just 7 months before magnets would explode at the LHC delaying its scientific mission by over a year, the anti-k_T algorithm was introduced [211]. Unlike k_T or C/A, anti-k_T tends to cluster hard, collinear cores first, and then soft radiation about that, and as such produces very regular, nearly perfectly disk shaped jet regions of radius R. Representative shapes of jets clustered with these algorithms are illustrated in Fig. 9 from Ref. [40]. Such jets have very consistent areas and are therefore easily calibrated experimentally, and before the LHC physics program started again in 2009, the anti-k_T algorithm was universally adopted as the jet finding algorithm of choice.

In concert with the development of the anti-k_T algorithm, at the same time (and even by some of the same authors!) another historical problem with pairwise recombination algorithms was solved. The naive computational complexity of a pairwise recombination algorithm is very expensive. For an event with N particles, we need to calculate the $\sim N^2$ pairwise metrics $d_{ij}^{(n)}$, and then we have to perform the clustering step $\sim N$ times to put every particle in a jet. Thus, a pairwise recombination algorithm costs $\sim N^3$, which is seriously slow and was a major sticking point for experiments to adopt it. This large time cost meant, for example, that jet finding could never be done in early triggering because there simply would not be enough time between bunch collisions for jets to be found. However, these authors noted that for particles beyond an angular scale of R, clustering proceeds independently, until large enough structures are produced. So, by

---

[9]My undergraduate research advisor, Steve Ellis, was for many years one of the very few voices in the wilderness pushing back on experimentalists' preferences for seeded cone algorithms, strongly advocating for widespread use of the IRC safe k_T algorithm, or even small modifications of the seeded cone algorithms that rendered them IRC safe [221]. Ellis is also perhaps most famous for a talk he gave in winter 1985 that countered Carlo Rubbia's claim that supersymmetry had been discovered in the UA1 and UA2 experiments [222]. This episode is documented in Gary Taubes's book *Nobel Dreams: Power, Deceit and the Ultimate Experiment.*
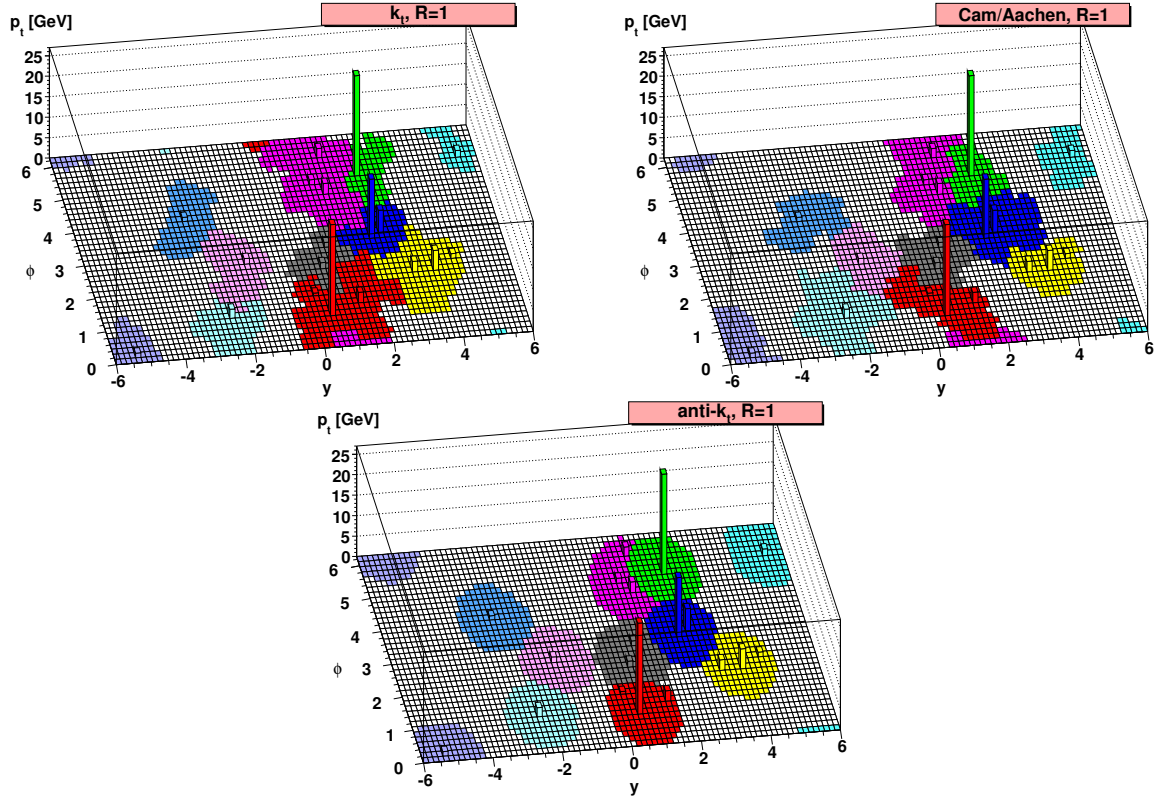
Figure 9: Demonstration of jet shape dependence from simulated collision events at the LHC. The k_T (top left), C/A (top right), and anti-k_T (bottom) algorithms are used to cluster particles in the same event. Extremely low energy particles, called "ghosts", are uniformly distributed over the event and distinct clustered jet regions are colored. Figure from Ref. [40].

tessellating the experiment into independent regions according to a fixed minimal distance between particles by the clustering metric (which itself is a fundamental problem in computational geometry [223, 224, 225, 226, 227, 228]), clustering could be dramatically more efficiently calculated, leading to huge gains in compute times, reducing the complexity to a mere $\sim N \log N$ scaling. This development led to the software program FastJet [229], which is universally used for jet finding and analysis by theorists and experimentalists alike.

3.4.2 Recombination Scheme

Unlike the clustering metric, there has in general been very little study of the recombination algorithm. At one level, it is "obvious" that one should use E-scheme recombination in which particle momenta are summed at every step of the clustering. This indeed has some nice properties, namely, that the jet's net momentum as a sum of momenta of its constituents is sequentially built up by the clustering, and, because momentum is conserved in the collinear splitting process that produced all those particles anyway, the net momentum of a jet can be interpreted as the momentum of the initiating parton. E-scheme recombination was advocated as a standard in a Les Houches

38

Accord [230], is the default setting in FastJet, and is what experiments use almost exclusively, though that need not be the case.

For its simplicity and ubiquity, E-scheme recombination does have some issues. Perhaps the biggest is its sensitivity to **recoil** from contamination radiation that just happens to land in the jet. Because jet radii must be finite in size, $R > 0$, there is some probability that radiation uncorrelated or only weakly correlated with the jet's dynamics will land in the jet's area. Such radiation can be produced as dipole radiation off other colored particles in the event, from the **underlying event**, secondary parton collisions in each proton collision at the LHC, or from **pile-up**, secondary proton collisions in each bunch crossing.[10] Even in the extreme case when there is a single hard particle of energy E and a single soft particle from contamination of energy E_s, E-scheme recombination will result in a jet direction that is displaced from the direction of the hard particle. Radiation uncorrelated with the jet will be approximately uniformly distributed over the jet's area, and further most of the area is near the edge of the jet, an angle R away. The angle between the direction of the hard particle and the E-scheme jet axis will then be

$$\Delta\theta \approx \frac{E_s}{E}\,R\,, \tag{3.49}$$

which may be significant, depending on the application.

It is therefore at the very least interesting to consider other recombination schemes, to see if another prescription might assuage the effects of recoil. The net momentum or E-scheme recombination axis of a jet is like the direction of mean momentum flow, and like the mean of a set of numbers, is very sensitive to outliers. Other averages are less sensitive to outliers, such as the median, so if there were a way to determine the "median" jet axis that might solve the problem. Actually, the median axis can be at least algorithmically defined [239], but is a computationally hard problem because jet finding proceeds in the two dimensions of the celestial sphere. (The median in one dimension is simple to evaluate, however.) Actually, determining the median jet axis is closely related to optimal transport theory [240, 241, 242, 243, 244, 245, 246, 247, 248] for jet physics in two dimensions, which has recently seen a burst of interest (see, e.g., Refs. [249, 250, 251, 252, 253]).[11]

Is there another way to proceed to define an approximate median jet axis, that is computationally simple yet is insensitive to recoil? Let's see if we can make sense of recombination schemes in general, in a similar way as the general form of the k_T-type metric. At every step of the clustering, the clustered particle must carry the total energy of its constituents, by collinear safety, but its direction can in principle lie anywhere between them. This ensures that in the exact collinear limit, the direction and energy of the clustered particle is exactly that of what you could in principle measure. So, one way to combine clustered momenta i and j into (ij) is as follows:

$$E_{(ij)} = E_i + E_j\,, \tag{3.50}$$

---

[10]A significant theoretical issue with the measurement of observables on a jet or other restricted region of phase space are **non-global logarithms** that arise from correlated emissions that land inside and outside of the region of interest [231]. The resummation of non-global logarithms is accomplished by a non-linear equation, the BMS equation [232], rather unfamiliar to usual linear renormalization group evolution that resums "global" logarithms. Several groups have developed techniques for systematic improvement of the description of non-global logarithms, see, e.g., Refs. [233, 234, 235, 236, 237, 238]

[11]For a modern mathematical treatise on optimal transport theory, see Ref. [254], by Cédric Villani. Villani won the Fields Medal in 2010 for his work on optimal transport theory [255, 256], and has recently been active in politics, having been elected and served in the French National Assembly. He is also strongly associated with wearing a distinctive spider brooch.

$$\hat{p}_{(ij)} = \frac{E_i^n \hat{p}_i + E_j^n \hat{p}_j}{\left| E_i^n \hat{p}_i + E_j^n \hat{p}_j \right|} \,. \tag{3.51}$$

The first line, the sum of energies is straightforward, and on the second line, the directions are weighted by powers of the energies of the particles (and then normalized to return the unit vector direction of the clustered particle). E-scheme recombination corresponds to $n = 1$ here, but infrared and collinear safety only limits us to consider $n > 0$. The E^2-scheme has been considered historically [208, 257], where one takes $n = 2$. This weights the direction of the clustered particle in such a way that it is much more strongly correlated with the harder particle, but there is still residual sensitivity to the softer particle.[12]

We aren't quite yet at a median, but let's keep going, and while we're at it, might as well crank n up to infinity. For $n \to \infty$, the direction of the clustered particle lies *exactly* on the direction of its harder constituent, and is correspondingly referred to as the **Winner-Take-All (WTA)** recombination scheme [259, 239, 260]:

$$\hat{p}_{(ij)}^{\text{WTA}} = \Theta(E_i - E_j)\, \hat{p}_i + \Theta(E_j - E_i)\, \hat{p}_j \,. \tag{3.52}$$

This simply ignores the softer of the two particles and therefore is almost even better than a median. Indeed, the resulting jet axis from WTA recombination is like the jet's **medioid** of momentum flow, the single particle that best represents the net momentum flow. Because at every stage of the clustering the WTA axis always lies on a particle, the WTA axis of the entire jet necessarily lies on one of the particles in the jet (rather unlike the E-scheme). The WTA recombination axis is therefore insensitive to recoil, solving that issue with the E-scheme. Unfortunately, the WTA axis introduces other issues or confusions that are not with the E-scheme, like the fact that at every stage of the clustering, the WTA scheme forbids a generation of non-zero mass, and instead every clustered particle must be treated as massless. Nevertheless, the WTA recombination scheme has proved very useful for applications of studying structure inside an already clustered jet, some applications of which we will touch on in the following section.

The effect of recoil on the direction of the reconstructed jet axis is displayed in Fig. 10, re-produced from a study in Ref. [261]. Here, jets have been simulated at a future 100 TeV hadron collider and clustered with the anti-k_T algorithm with a radius $R = 0.5$. In this sample of events, contamination radiation has been included through simulation of additional proton collisions per bunch crossing, defined by the number of primary vertices, N_{PV}. Then, the angle between the original jet axis, with no additional proton collisions, is compared to the jet axis after including this contamination radiation. With E-scheme recombination, as the number of primary vertices increases, the distribution of angles extends to larger and larger values, with a long tail even out to the jet radius itself. By contrast, jets constructed with WTA recombination have nearly identical axes regardless of the amount of contamination radiation.

For their enormous practical importance, jet algorithms will play only a minor role in the rest of these lectures. We will only need the consequence of IRC safety that collinear radiation is always captured in a single jet, regardless of the specific algorithm, so we will always assume that we have already clustered and found jets, and are then looking deep inside of them.

---

[12]Mike Seymour used to have a website with his homebrewed software that included an implementation of the k_T algorithm and allowed for changing the recombination scheme, as well EVENT2 [258], but the site seems to have dissolved. I don't know if the site still exists but hosted elsewhere.
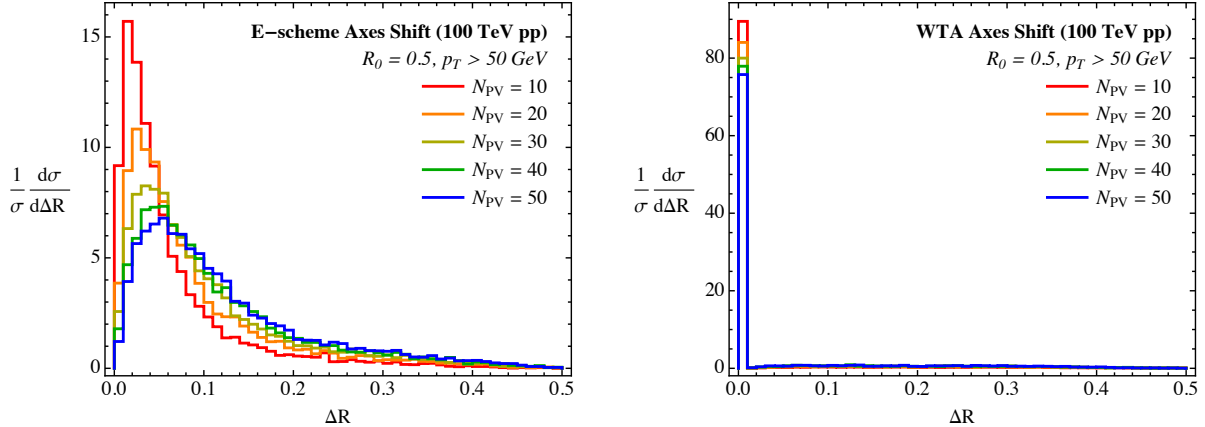
Figure 10: Plots of the distribution of angular shift ΔR in the jet axis as the amount of contamination radiation in a jet is increased. Jets are clustered with the anti-k_T algorithm with a radius $R = 0.5$ with a minimum transverse momentum of 50 GeV in simulated data from a future 100 TeV hadron collider. The number of primary vertices N_{PV}, corresponding to the number of simulated proton collisions per bunch crossing, is a proxy for increased contamination. Left: Jet axis shift with E-scheme recombination. Right: Jet axis shift with WTA recombination. From Ref. [261].

Exercises

1. Express three-body collinear phase space in terms of three energy fractions z_1, z_2, z_3, two angles between pairs of particles θ_{12}, θ_{13}, and one azimuthal angle ϕ. You should find [262, 263]

$$d\Phi_{\mathrm{coll}}^{(N=3)} = \frac{4E^4}{(4\pi)^5} \, d\theta_{12}^2 \, d\theta_{13}^2 \, d\phi \, dz_1 \, dz_2 \, dz_3 \, z_1 \, z_2 \, z_3 \, \delta(1 - z_1 - z_2 - z_3) \,. \tag{3.53}$$

2. Calculate the distribution of the angle between the harder particle and the net momentum axis in a two-particle quark jet. Assume that the two particles in the jet are within an angle R of each other. Now, calculate the distribution of the angle between the harder particle and the general E^n-scheme axis, as a function of n. See Ref. [264] for more details.

3. On flat phase space in the center-of-mass frame with total energy Q, determine the single-particle energy distribution, $p(E)$. Take the large number of particles, $N \to \infty$, limit; what does the distribution reduce to?

4 First Example: Quark vs. Gluon Jet Discrimination

With our introductions to the necessary tools of machine learning and perturbative quantum field theory, we now turn our focus on using those results in practice, to understand a number of problems of central importance for jet substructure. We start in this section, and will often return back later, with the problem of discrimination of jets initiated by quarks versus those initiated by gluons. This problem is essentially as old as QCD itself, through the discovery of the gluon through the observation of three-jet events in e^+e^- collisions [33, 34, 35, 36, 37]. Quarks are electrically

charged, and so can be probed by deeply-inelastic electron scattering on a proton, but the gluon is electrically neutral and so must be produced directly from the quarks as color-charged particles themselves. The validity of QCD as *the* theory of the strong force requires the existence of the gluon so its unambiguous observation is clearly of paramount importance.[13]

4.1 Some History and Context

As stated, this is a binary discrimination problem, we just want to distinguish quark jets (background) from gluon jets (signal) through a detailed study of the internal dynamics of the jets. Beyond validation of QCD, a powerful quark versus gluon discriminator would have profound impact in nearly the entire particle collider physics program. Its importance (and, as we will see, its challenge), then, has prompted some to refer to this as the "white whale of jet substructure",[14] but I hope that this ominous description just implies that our quarry is worth the effort and not the associated doom of its literary inspiration.

Regardless of your interests, a quark versus gluon discriminator can help you. If you are interested in parton distribution function (pdf) extractions, quark pdfs are accurately determined from deeply-inelastic scattering data or Drell-Yan production, but direct sensitivity to the gluon pdf requires final-state jet (or Higgs) production at the LHC [265, 266]. If you are interested in final state fragmentation, fragmentation functions are accurately measured on jets in e^+e^- collisions, but the vast majority of those jets are initiated by quarks, and gluon jets at high energies are only produced in significant numbers at a hadron collider [267, 268, 269, 270, 271, 272]. In a related vein, if you are interested in tuning Monte Carlo event generators [273, 274, 275, 276, 277, 278, 279], e^+e^- collision data strongly constrains quark jets, but without a pure sample of gluon jets, data from hadron colliders is much more challenging to interpret. If you are interested in physics beyond the Standard Model, cascade decay chains initiated by new, unstable particles can often produce huge numbers of quark jets in a single event, e.g., Refs. [280, 281, 282, 283, 284, 285, 286, 287], while background Standard Model processes would dominantly produce gluon jets copiously. And then, if you are like me and are just interested in the substructure of jets on its own, you just want to determine the method for most efficient extraction of the type of jet simply from the measurable momenta of their constituent particles.

Quark versus gluon discrimination as a problem in its own right has a very long history in particle physics, and in fact some of the first neural networks applied to analyze simulated collider events were employed to this task [288, 289, 290, 291, 292, 293, 294, 295]. With the rise of jet substructure, early work in the LHC era was focused around designing bespoke observables for quark versus gluon discrimination and combining and slicing high-dimensional observable spaces just to wrap one's head around where on this space different types of jets lived [296, 297, 298]. More and more techniques were subsequently developed, and the community took up the problem and worked to formalize it in a Les Houches study [299]. Since that time, more and more of the discrimination task has been pushed onto deep neural networks, with lower and lower level data as the input.

However, while machine learning discrimination performance is extremely impressive, we want to understand why it is and what physics of quark and gluon jets are responsible for the differences. This is our task in these lectures and is a vital handrail and guide for machine learning studies

---

[13]The gluon, however, is one of the few elementary particle discoveries for which there was no associated Nobel Prize.

[14]I believe this phrase is credited to my post-doc advisor Jesse Thaler.

What is a Quark Jet?
From lunch/dinner discussions

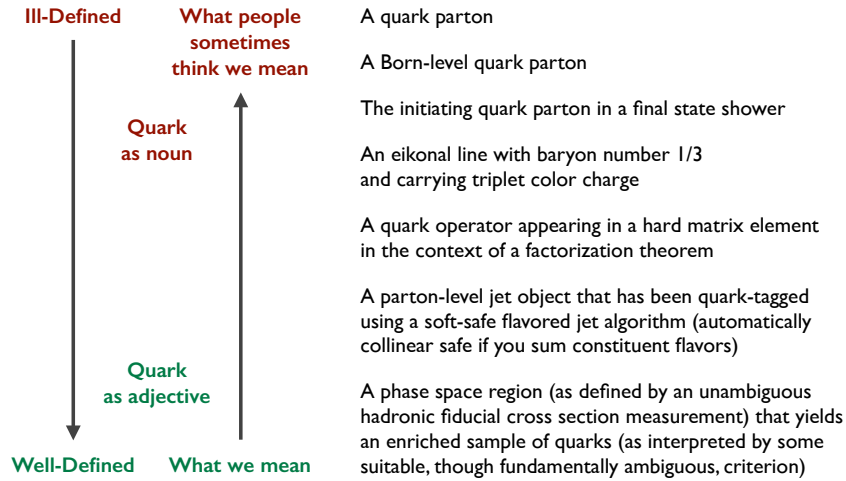|		
Ill-Defined	**What people sometimes think we mean**	A quark parton
		A Born-level quark parton
Quark as noun		The initiating quark parton in a final state shower
		An eikonal line with baryon number 1/3 and carrying triplet color charge
		A quark operator appearing in a hard matrix element in the context of a factorization theorem
		A parton-level jet object that has been quark-tagged using a soft-safe flavored jet algorithm (automatically collinear safe if you sum constituent flavors)
Quark as adjective		A phase space region (as defined by an unambiguous hadronic fiducial cross section measurement) that yields an enriched sample of quarks (as interpreted by some suitable, though fundamentally ambiguous, criterion)
Well-Defined	**What we mean**	

Figure 11: Summary of thoughts and discussions of a more robust and physical definition of a "quark jet" from the 2015 Les Houches study. From Ref. [299].
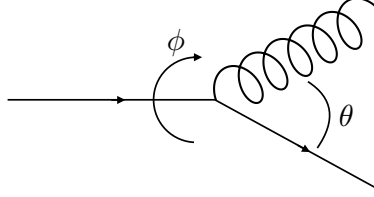
because they nearly exclusively are trained on simulated data from an event generator simulation program. While at their core event generators use fixed-order matrix elements or DGLAP evolution, there are still many tuned parameters and fits and it may be that a machine is merely learning idiosyncratic differences of quark and gluon jets from the way they are simulated in a particular program, and not actually some fundamental physical distinction. This is especially relevant as mentioned above because a pure sample of gluon jets in data is challenging to produce and it has been noted that while quark jets look more or less the same in different event generators, gluon jets can vary significantly [299]. So, part of our goal here will also be to determine those features that are robust, that can inform better, more principled, machine learning, even when the simulated data may be noisy.

A bigger problem with a quark versus gluon discriminant endeavor is even more basic. As discussed in the Neyman-Pearson lemma proof, of course the likelihood ratio is the optimal discriminant. However, the likelihood ratio requires perfect ground truth of what "signal" and "background" are; we condition on the label "signal" and "background", and then measure other quantities on these samples. Jets at the LHC obviously do not come with labels on them (or our job would be very easy!), but jets in simulation do, and this can lead to confusion or at least imprecision. In an event simulation, you tell the simulator what process to generate, typically defined at leading-order in the strong coupling. In practice, this "Monte Carlo truth label" is then used to define the ground truth and corresponding conditioned distributions of signal and background: if you requested events with a final state gluon, then the resulting event contains a gluon jet. While the type or flavor of a jet is well-defined at leading order, such a naive label does not generalize to higher orders [300]. This point was identified in the community study of Ref. [299], and there, worked to broaden how we refer to jet types to have a more physical foundation. The outcome of

informal discussions during the 2015 Les Houches study on more robust definitions of a "quark jet" is presented in Fig. 11. Nevertheless, Monte Carlo truth labeling remains the common practice, and so another aspect of our work here will be to establish when and where such a label is meaningful and accurate.

4.2 Simplicity of Jets at High Energies

We will start our analysis of quark versus gluon jet discrimination as simple as possible, but no simpler. We could use our master formula derived in the previous section to calculate distributions from the known splitting functions, but I want to take a different route here that clearly illustrates the assumptions we employ.[15] We will thus ignore the details of QCD, ignore Feynman diagrams, etc., and, like ignorant scientists taking the first great leap into the unknown, simply make some hypotheses and follow them to the edge of logic. These will be so important that we will refer to them as axioms.

For our first foray into quark versus gluon discrimination, our axioms will be:

1. At high energies, QCD is an approximately scale-free, conformal field theory. This means that, to first approximation, the coupling of QCD, α_s, is constant, independent of energy.

2. The coupling constant of QCD, α_s, is small, $\alpha_s \ll 1$, and so quarks and gluons are good quasi-particles.

3. The only difference between quarks and gluons are their color charges C_F and C_A, respectively, the quadratic Casimirs of the fundamental and adjoint representations of SU(3) color, with $C_A > C_F$.

With these axioms, we will ultimately establish baseline discrimination performance for this problem, calculate the ROC curve, and all that, but let's first start much simpler and just attempt to calculate the rate or probability that a quark emits a gluon in this framework, $P_{q \to qg}$. In the context of Feynman diagrams, what we would want to compute is

$$
P_{q \to qg} = \left| \vphantom{\Bigg|} \right|^2 , \tag{4.1}
$$

but we don't have Feynman diagrams, so we need to use our axioms alone.

First things first: let's determine what possible quantities this probability can depend on. Specifically, we want to determine this probability as a function of the properties of the emitted gluon. So, what possible properties can the gluon have? If we only consider measuring momentum, then the four vector of the gluon p_g completely specifies it. Let's choose some convenient coordinates to express the four vector. In particular, there is no notion of absolute momentum in a Lorentz-invariant theory, so we can express the gluon's four vector in relationship to the final state quark.

---

[15]The approach especially of this lecture is strongly influenced by TASI 2001 lecture notes by my undergraduate research advisor, Matt Strassler [301].

Further, the gluon is massless (by the scale invariant assumption), and so its four-vector can be expressed as

$$p_g = E_g(1, \sin\theta\cos\phi, \sin\theta\sin\phi, \cos\theta) \,. \tag{4.2}$$

Here, θ is the angle between the quark and the gluon and ϕ is the azimuthal angle of the gluon about the quark:

However, with no preferred spatial orientation, we can freely set the azimuthal angle to 0 with no physical consequence (there is an SO(2) $\simeq$ U(1) symmetry about the initial quark).

We can express the gluon's energy E_g in a convenient way, too. The initial quark has some fixed energy E set by the experimental set-up, and the gluon energy is some fraction z of that: $E_g = zE$, where $z \in [0,1]$. With these identifications, the four-vector can be expressed as

$$p_g = zE(1, \sin\theta, 0, \cos\theta) \,, \tag{4.3}$$

and this makes it clear that there are two degrees of freedom of the gluon: the angle θ and the energy fraction z. We can then write the original probability that we want as

$$P_{q\to qg} = p(z,\theta)\, dz\, d\theta \,, \tag{4.4}$$

where $p(z,\theta)$ is a probability density to be determined.

Let's see what we can do about this probability density. "Scale invariant" means that the physics is unchanged if you change the scale at which you view the system. That is, physical quantities are unaffected by a scaling, and a physical quantity is indeed the rate of gluon emission. So, somehow this probability for gluon emission must be unchanged by a scale transformation. One possible scale transformation is to modify the splitting angle by some $\lambda > 0$, as $\theta \to \lambda\theta$. Let's write this as an infinitesimal transformation with $\lambda = e^\epsilon$, for $\epsilon \ll 1$. Then, under this transformation, the measure becomes

$$dz\, d\theta \to dz\, d\theta + \epsilon\, dz\, d\theta + \cdots \,, \tag{4.5}$$

suppressing terms higher order in ϵ. The probability density, by contrast, transforms as

$$p(z,\theta) \to p(z, e^\epsilon\theta) = p(z,\theta) + \epsilon\theta\frac{d}{d\theta}p(z,\theta) + \cdots \,. \tag{4.6}$$

If the probability is to be invariant under this transformation, then combining these results we must have

$$P_{q\to qg} \to p(z, e^\epsilon\theta)\, dz\, d(e^\epsilon\theta) = \left(p(z,\theta) + \epsilon\theta\frac{d}{d\theta}p(z,\theta) + \cdots\right)(dz\, d\theta + \epsilon\, dz\, d\theta + \cdots) \tag{4.7}$$

45

$$= p(z, \theta) \, dz \, d\theta \, .$$

Now, we expand in ϵ, and cancel terms. The ϵ^0 term explicitly cancels with the untransformed probability on the second line, while at ϵ^1, we have the requirement that

$$\theta \frac{dp(z, \theta)}{d\theta} + p(z, \theta) = 0 \, , \tag{4.8}$$

tossing away the benign differential elements. The solution of course is that

$$p(z, \theta) \propto \frac{1}{\theta} \, . \tag{4.9}$$

The exact same argument goes through with scaling the energy fraction z as well, and one finds the same functional form for the z dependence. We can, as we did here, independently scale energies or angles and so the only possible solution consistent with scale invariance that allows for independent scalings is

$$p(z, \theta) \, dz \, d\theta \propto \frac{dz}{z} \frac{d\theta}{\theta} \, . \tag{4.10}$$

This result is rather interesting, and one thing to note at this stage is that demanding scale invariance highly constrains the gluon's dynamics. The only possible way that you can have a valid scale transformation of the energy fraction $z \in [0, 1]$ and maintain energy conservation is if the gluon's energy was parametrically smaller than the quark's energy, $z \lll 1$. Similarly, if there was any constraint on the emission angle of the gluon, like it must lie in a jet of radius R, scale invariance enforces that the emission angle is also parametrically small, $\theta \lll R$. Thus, we see that we are forced into the soft and collinear regime by our axioms.

Now, we turn to our second and third axioms, which set the scale of the probability density. The coupling α_s determines the strength with which the gluon couples to the quark, and the quadratic Casimir C_F quantifies how the quark and gluon share color. Thus, the probability density is proportional to both, $p(z, \theta) \propto \alpha_s C_F$. We'll now cheat a tiny bit and use our knowledge of the soft and collinear limit of the splitting functions from Sec. 3.2.1, and get all of the factors of 2 and π correct:

$$P_{q \to qg} = p(z, \theta) \, dz \, d\theta = \frac{2\alpha_s C_F}{\pi} \frac{dz}{z} \frac{d\theta}{\theta} \tag{4.11}$$

Of course, this is exactly what we had written down earlier for the soft and collinear limit of gluon emission off of a quark, but now we see that it follows from the strong constraint of scale invariance.

4.2.1 The Lund Plane

With this result, we can now attempt to interpret it. The first thing to notice, and you might have for some time already, is that this is not a probability. By scale invariance, we are forced into the regime where $z, \theta \to 0$, but the probability density is non-integrable there

$$\int_0 dz \int_0 d\theta \, p(z, \theta) = \frac{2\alpha_s C_F}{\pi} \int_0 \frac{dz}{z} \int_0 \frac{d\theta}{\theta} \to \infty \, . \tag{4.12}$$

Lacking integrability, $P_{q \to qg}$ is also not a probability, and so we seem to have lost all of the gains that the axioms had earned us. However, let's think a bit more critically about what happens in

the problematic limits where this purported probability density is non-integrable. Consider the $z \to 0$ limit, corresponding to the emission of a 0 energy gluon. Such a gluon is, by definition, not observable because no experiment you can make can detect exactly 0 energy. Additionally, strictly 0 energy gluons carry, well, no energy, and so we can emit an arbitrary number of them without any effect whatsoever on any observable quantity. So, the problem with our probability so far is not really that it diverges, it is that we haven't fully accounted for what it describes. Because no measurement can distinguish 1 emitted 0 energy gluon from 143,872 emitted 0 energy gluons, we need to sum up all possible numbers of emitted 0 energy gluons, by the rules of quantum mechanics.

An analogous argument holds for a 0 angle emission. No experiment can identify a gluon emitted exactly collinearly with a quark, and so again, we need to sum up all possible numbers of exactly collinear gluon emissions, by the rules of quantum mechanics. Actually, now is where we see the KLN theorem arise. We had been studying a system in degenerate perturbation theory (single soft-collinear gluon emission), and if we do that, infinities blow up in our face. As we will demonstrate shortly, by summing up all possible degenerate states that have exactly the same experimental signatures but different numbers of gluons, we will find finite quantities and true probability distributions, just as KLN promised.

Let's see how to do this. Back to our emission "probability", we can rewrite it as

$$P_{q \to qg} = \frac{2\alpha_s C_F}{\pi} \frac{dz}{z} \frac{d\theta}{\theta} = \frac{2\alpha_s C_F}{\pi} d \left(\log \frac{1}{z} \right) d \left(\log \frac{1}{\theta} \right) , \qquad (4.13)$$

which is now flat logarithmically in energy fraction and angle. This now provides a justification and interpretation of what is going on and how to fix it all up. Where this probability is small, specifically $P_{q \to qg} < 1$, this is indeed sensible as a probability. However, where it gets large, it no longer should be interpreted as a probability, but rather as an *expectation value* of the number of gluons emitted with energy fractions and splitting angles less than given values (i.e., less resolvable than some values of z, θ).[16] The uniform distribution renders this interpretation immediate, because an integral over a uniform distribution just returns the upper bound. Therefore, we nicely visualize a large number of gluons emitted according to this uniform distribution.

The visualization of this is shown at left in Fig. 12 in which phase space of gluon emissions is represented as a semi-infinite plane in $\log \frac{1}{z}$, $\log \frac{1}{\theta}$ and each dot represents an emitted gluon with that value of energy fraction and angle. This representation is called the **Lund plane** [305] (see also Ref. [306] and recent work and measurements of Refs. [307, 308, 309, 310, 311, 312, 313]) after the researchers in Sweden that introduced it. This uniform emission plane is the starting point for particle production in a modern parton shower Monte Carlo generator, like Pythia, Herwig, or Sherpa [314, 315, 316, 317], though they include much, much more physics, like the full splitting functions, the running coupling (which increases emission rate further from the origin), or hadronization (which cuts off perturbative emissions at some scale).[17] Those pesky divergences are

---

[16]This was likely known or at least appreciated long before (perhaps first by Wigner [302]), but Dirac [303] and Feynman [304] noted that predictions from quantum field theory of probabilities that are larger than 1 (as in the case at hand), must be associated with probabilities that are negative, so that their sum produces physical probabilities that lie on $[0, 1]$. Further, such negative probabilities simply must make sense on their own.

[17]There has been significant work recently in developing parton showers that are formally accurate to much higher orders than these workhorse generators, e.g., Refs. [318, 319, 320, 321, 322]. A particularly important point of these parton showers, and for their generalization to even higher orders, is an **ordering variable**, which is an observable quantity that monotonically increases across the Lund plane and ensures that emissions are produced in such a way to correctly incorporate all of the logarithmic structure to the accuracy one is working.
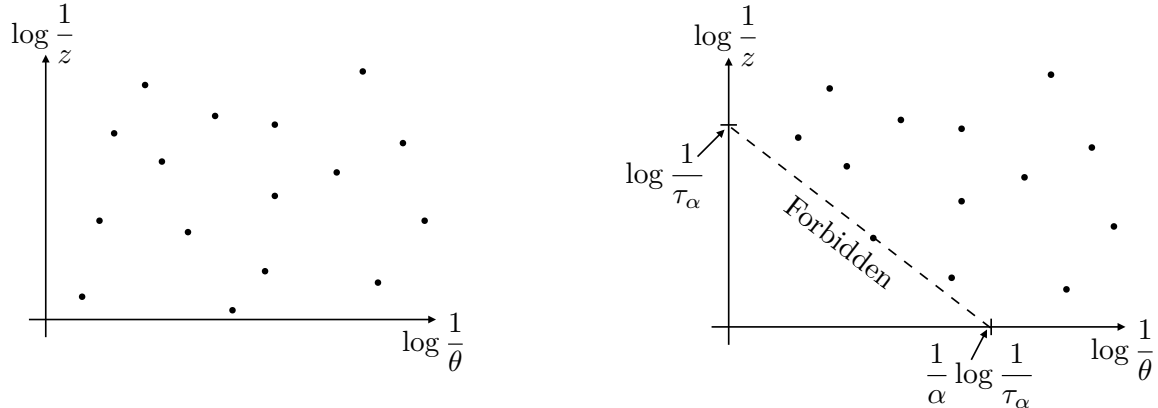
Figure 12: Left: Visualization of emissions uniformly populating the Lund plane. Right: A measured value of the angularity τ_α enforces no emissions below the line of constant angularity.

pushed off to infinity, the soft divergence vertically and the collinear divergence horizontally, and the only vestiges of divergences that remain is the fact that this Lund plane, as we have constructed here, is infinite in area. Once we start asking reasonable, IRC safe questions, about such jets on the Lund plane, however, divergences will be no where in sight.

4.2.2 Resummation and the Sudakov Form Factor

The Lund plane captures the distribution of soft and collinear emissions off of a hard initiator particle and somehow this distribution must encode properties of that initiator particle. Our goal, for quark versus gluon jet initiator discrimination, is to find out how that information is encoded. As a first baby step in that direction, we will just calculate one of the oldest jet substructure observables called the **angularities** [323, 324, 325]. The angularities measure the net flow of energy and momentum away from the jet axis or initial hard particle and is defined as

$$\tau_\alpha = \sum_i z_i \theta_i^\alpha \,. \tag{4.14}$$

The sum runs over all particles in the jet, and the parameter $\alpha > 0$ ensures IRC safety. Note that this is linear in the energy fractions z_i, which is a requirement of collinear safety to ensure that the energies of exactly collinear emissions sum correctly. The angle θ_i is measured from particle i to the jet axis, and here, because all emissions are soft and collinear, the jet axis coincides with the initial hard particle. Thus, the only particles that contribute and have a non-zero angle from the jet axis are the emitted gluons. For the rest of this section, we will determine how the angularities can be used for quark versus gluon discrimination.

Angularities, in one form or another, have been studied for over 50 years in QCD, with the $\alpha = 2$ case corresponding to thrust [326, 327]. Thrust is one of the most studied, calculated, and experimentally measured jet observables, because, in the collinear limit, thrust is proportional to the jet mass, which is perhaps the most basic jet substructure observable whatsoever [328]. Another ancient angularity is the broadening [329, 330], corresponding to $\alpha = 1$ exponent. Broadening has the interesting feature (or bug) that it is sensitive to recoil [331, 332], and correspondingly

48

is sensitive to the precise jet axis definition.[18] We will discuss this property a bit more later when we discuss the effect of higher orders on discrimination power. Observables that are related to angularities are energy correlation functions [334, 335, 336, 298, 337] which quantify average pairwise radiation patterns. Moments of energy correlation functions are energy-energy correlators [338, 339, 340, 341, 342, 343, 344, 345] introduced long ago, but are becoming a very hot topic because their simplicity enables extremely precise theoretical predictions.

But back to angularities for now. The first thing we need to do to determine the probability distribution of the angularity τ_α is to evaluate it on our jet of interest. Emissions on the Lund plane are uniformly distributed in the logarithm of the splitting angle and the energy fraction. Therefore, in "real" space, linear in splitting angle and energy fraction, emissions are exponentially far apart. This observation makes evaluation of the angularity very simple. There will be one emission in the jet for which $z_i \theta_i^\alpha$ is the largest, and then all other emissions' contributions to the angularity will be exponentially small. So, to exponential accuracy, we can consider the value of the angularity set by that one emission alone. This is illustrated at right in Fig. 12.

Then, with this one emission, note that a fixed value of the angularity defines a line on the Lund plane:

$$\log \frac{1}{\tau_\alpha} = \log \frac{1}{z} + \alpha \log \frac{1}{\theta} \,, \tag{4.15}$$

which is dashed at right in Fig. 12. If τ_α is indeed to be the measured value of the angularity on a jet, then there must be no emissions below that dashed line, toward the origin of the Lund plane. Emissions below that dashed line correspond to larger values of the angularity and are forbidden by our measurement. However, emissions above the dashed line are completely unconstrained, by our assumption that they do not contribute to the angularity anyway.

Then, to calculate the probability distribution of the angularity, we just need to calculate the probability that there are no emissions below that dashed line whatsoever. To do this, we will break up the forbidden triangle region into many small pieces:
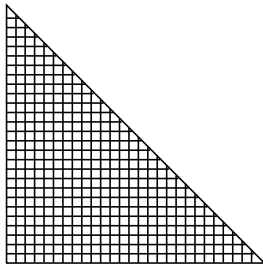
There can be no emissions in any of the subregions of the triangle, which is an "and" statement in probability and so we must multiply all of the individual no emission probabilities together. If there are N equal area subregions, then the probability for emission in any subregion is equal and is

$$\text{Probability for emission} = \frac{2\alpha_s}{\pi} C_F \frac{\triangle}{N} \,, \tag{4.16}$$

---

[18]Spherocity [333] is a broadening-like observable that was introduced shortly after thrust that defines the jet or event axis as the direction that minimizes the sum of magnitudes of momentum transverse to it. As such it is actually insensitive to recoil. We will discuss such "broadening axes" a bit more later with higher-order effects.

where $\triangle$ is the area of the forbidden triangle:

$$\triangle = \frac{1}{2\alpha} \log^2 \tau_\alpha \, . \tag{4.17}$$

This follows because emissions are uniformly distributed on the Lund plane and so probabilities are proportional to the area of the region of interest.

The probability of no emission in a small subregion is simply 1 minus the emission probability (either a particle is emitted or not):

$$\text{Probability for no emission} = 1 - \frac{2\alpha_s}{\pi} C_F \frac{\triangle}{N} = 1 - \frac{\alpha_s}{\pi} \frac{C_F}{\alpha} \frac{\log^2 \tau_\alpha}{N} \, . \tag{4.18}$$

Now, we just need to multiply this no emission probability by itself N times to determine the no emission probability anywhere in the triangle. Taking the $N \to \infty$ limit to make the subregions infinitesimally small, we have

$$P(\tau_\alpha \text{ less than measured value}) = \lim_{N \to \infty} \left(1 - \frac{\alpha_s}{\pi} \frac{C_F}{\alpha} \frac{\log^2 \tau_\alpha}{N} \right)^N = \exp\left[-\frac{\alpha_s}{\pi} \frac{C_F}{\alpha} \log^2 \tau_\alpha \right] \, . \tag{4.19}$$

The product transmogrifies into an exponential! This exponential factor is called the Sudakov form factor [346], and is simply a manifestation of the scale-invariant Poisson process of particle emission in high-energy QCD. Note that we have called this the probability that τ_α is less than the measured value because smaller τ_α means that the leading emission is farther from the origin of the Lund plane, somewhere, anywhere above that dashed line.

Therefore, this probability is actually a cumulative probability distribution of the angularity of a quark jet,

$$\Sigma_q(\tau_\alpha) = \exp\left[-\frac{\alpha_s}{\pi} \frac{C_F}{\alpha} \log^2 \tau_\alpha \right] \, . \tag{4.20}$$

The differential probability, the probability distribution, is the derivative of this

$$p(\tau_\alpha) = \frac{d}{d\tau_\alpha} \exp\left[-\frac{\alpha_s}{\pi} \frac{C_F}{\alpha} \log^2 \tau_\alpha \right] = -\frac{2\alpha_s}{\pi} \frac{C_F}{\alpha} \frac{\log \tau_\alpha}{\tau_\alpha} \exp\left[-\frac{\alpha_s}{\pi} \frac{C_F}{\alpha} \log^2 \tau_\alpha \right] \, , \tag{4.21}$$

and is normalized on $\tau_\alpha \in [0, 1]$. Note that these expressions are Taylor series in α_s to all orders. Each emission costs α_s, so this exponential Sudakov factor indeed requires explicit summation over all degenerate states, with any and all numbers of emitted gluons, to render the distribution finite and integrable. In fact, this "summation" over degenerate states is referred to as **resummation** and is a necessary theoretical tool for precision control of the soft and collinear regions of phase space.

The particular resummation we have performed here is the lowest, non-trivial accuracy, referred to as **double-logarithmic accuracy**. It is double logarithmic in the sense that there are two factors of logarithm for every factor of α_s in the exponent. More generally, observables like angularities that are both IRC safe and additive (expressed as an explicit sum over contributions from each particle) enjoy a resummed cumulative distribution of the form [347]

$$\Sigma(L) = \exp\left[L\, g_1(\alpha_s L) + g_2(\alpha_s L) + \alpha_s g_3(\alpha_s L) + \cdots \right] \, , \tag{4.22}$$

where $L = -\log\tau_\alpha$. The functions in the exponent are ordered in their logarithmic accuracy, with $g_1(\alpha_s L)$ the leading-logarithmic term, $g_2(\alpha_s L)$ the next-to-leading logarithmic term, etc. Resummation formally treats the region of phase space where $\alpha_s L \sim 1$ as $\alpha_s \to 0$, so that the importance of the terms in the exponent are controlled by the explicit powers of α_s. The double-logarithmic term we calculated is the lowest-order part of the leading logarithmic function, $g_1(\alpha_s L)$.

4.2.3 Casimir Scaling and a Discrimination Baseline

We could do this whole rigmarole again for a gluon jet, but by axiom #3, we know that the only difference would be in the replacement of the color factor C_F with C_A everywhere. So, let's just do that directly to the cumulative distribution:

$$\Sigma_g(\tau_\alpha) = \exp\left[-\frac{\alpha_s}{\pi}\frac{C_A}{\alpha}\log^2\tau_\alpha\right].$$
(4.23)

A plot comparing these double logarithmic distributions for angularities measured on quark and gluon jets is displayed at left in Fig. 13. This immediately shows us that the quark and gluon cumulative distributions satisfy the scaling relationship

$$\Sigma_g = \Sigma_q^{C_A/C_F}.$$
(4.24)

This is referred to as **Casimir scaling** [348, 349, 350, 351, 352], because quark and gluon jets only differ in their color Casimir factors and further, the dependence on the Casimirs is exponential.[19]

Now we're cooking because if we remember from very early on, this cumulative distribution relationship is precisely the ROC curve. Then, in the soft and collinear limit, where quarks and gluons only differ by their color charge, the ROC curve of measuring an angularity τ_α is

$$\text{ROC}(x) = x^{C_A/C_F} = x^{9/4}.$$
(4.25)

As a baseline for discrimination power, let's consider a couple of metrics. First, the ROC curve at $x = 0.5$, where 50% of the quark jets are kept, then

$$\text{ROC}(0.5) \approx 0.21,$$
(4.26)

21% of the gluon jets are also kept. Correspondingly, the fraction x of quark jets at which 50% of the gluon jets are kept, where $\text{ROC}(x) = 0.5$ is

$$1 - \text{ROC}^{-1}(0.5) \approx 0.27,$$
(4.27)

so about 27% of quark jets are kept. Finally, we can calculate the AUC,

$$\text{AUC} = \int_0^1 dx\, x^{C_A/C_F} = \frac{C_F}{C_F + C_A} \approx 0.31.$$
(4.28)

---

[19]Casimir scaling typically refers to the $C_F \leftrightarrow C_A$ relationship between the cusp anomalous dimensions Γ of the quark and gluon form factors. The one-loop cusp anomalous dimension is the coefficient of the logarithm in the exponent of the Sudakov factor, with $\Gamma_q = \frac{C_F}{C_A}\Gamma_g = \frac{\alpha_s}{\pi}C_F + \cdots$. Quadratic color Casimir scaling is known to be violated starting at four-loop order, when quartic color invariants first appear [353, 354, 355].
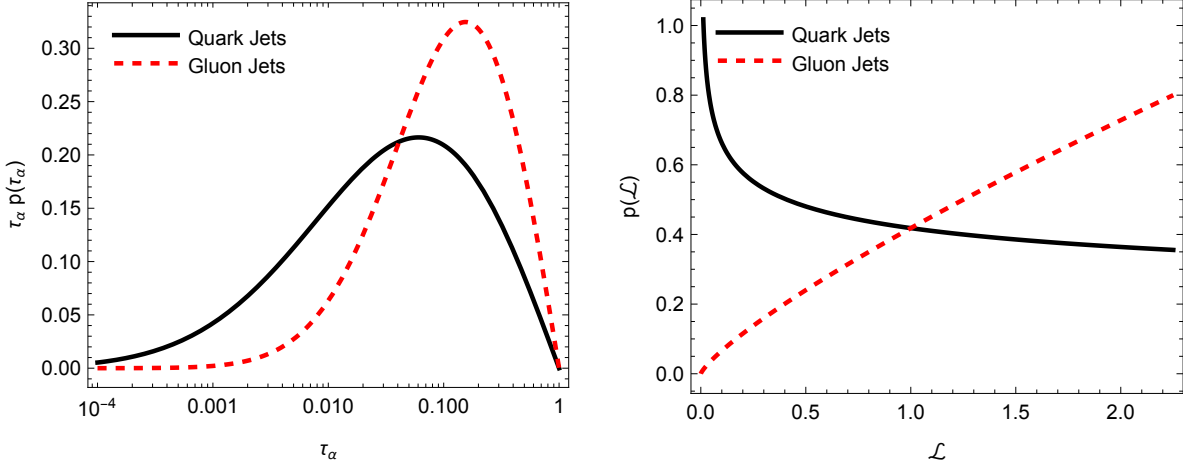
Figure 13: Left: Plots of the probability distributions of the angularity τ_α as measured on quark (black) and gluon (red) jets from the double logarithmic picture of uniform emissions on the Lund plane, Eq. (4.21) for quarks (and the corresponding distribution for gluons). Right: Distribution of the likelihood ratio $\mathcal{L}$, Eq. (4.29), as measured on quark (black) and gluon (red) jets from measurement of the angularity τ_α.

We can also calculate the reducibility factors corresponding to maximal attainable sample purity. In this case, the likelihood ratio is

$$\mathcal{L} = \frac{p_g(\tau_\alpha)}{p_q(\tau_\alpha)} = \frac{C_A}{C_F} \exp\left[-\frac{\alpha_s}{\pi} \frac{(C_A - C_F)}{\alpha} \log^2 \tau_\alpha \right] . \tag{4.29}$$

The distribution of this observable on quark and gluon jets is plotted at right in Fig. 13. As $\tau_\alpha \to 0$, the exponential factor vanishes because $C_A > C_F$, and so the minimum value of the likelihood is $\mathcal{L}_{\min} = 0$. That is, a pure sample of quark jets can be isolated by restricting the value of τ_α to be sufficiently small. By contrast, as $\tau_\alpha \to 1$, the likelihood ratio approaches its maximum of

$$\mathcal{L}_{\max} = \frac{C_A}{C_F} . \tag{4.30}$$

This means that, by taking $\tau_\alpha \to 1$, you can create a sample for which gluon jets and quark jets exist in the ratio $C_A : C_F$. If you only measure the angularity τ_α, you cannot isolate pure gluon jets with no quark jet contamination.

As yet another measure of discrimination power, we can quantify the **significance**, or the ratio of signal to the expected statistical fluctuation in the background. This metric is relevant for new particle discovery, and was, for example, used to quantify the 5σ discovery of the Higgs boson in 2012 [356, 357]. If we believe quantum mechanics, and collisions are truly independent, then events at the LHC are i.i.d., and the distribution of the number of signal or background events (i.e., the sum of all events satisfying a classification criteria) is governed by the central limit theorem. Therefore, the expected statistical fluctuation of the background events, the standard deviation of the number of background events, scales like the square-root of the number of background events. Then, the significance $\mathcal{S}$ is the ratio of the number of signal events S to the square-root of the

number of background events B:

$$\mathcal{S} = \frac{S}{\sqrt{B}} \,. \tag{4.31}$$

For the discovery of the Higgs, the significance was about 5, with several thousands of background events measured in the bins around 125 GeV, with an excess observed of only a few hundred events.

The significance can be massaged into a form where its relationship to the ROC can be made apparent. Let's denote the total number of signal (background) events in our sample as N_s (N_b). Now, given the likelihood ratio discriminant $\mathcal{L}$, we would like to make a cut on $\mathcal{L}$ and determine the significance with that additional cut. The fraction of signal (background) events that pass this cut is $1 - \Sigma_s(\mathcal{L})$ $(1 - \Sigma_b(\mathcal{L}))$, the corresponding cumulative distributions of the likelihood. Note that we subtract the cumulative distributions from 1 because (by our definition of the likelihood ratio), signal events accumulate at large likelihood, and so the cut value $\mathcal{L}_{\text{cut}}$ is the minimal value of the likelihood. Then, the significance with a cut on the likelihood ratio is

$$\mathcal{S}(\mathcal{L}) = \frac{N_s}{\sqrt{N_b}} \frac{1 - \Sigma_s(\mathcal{L})}{\sqrt{1 - \Sigma_b(\mathcal{L})}} \,. \tag{4.32}$$

The ratio $N_s/\sqrt{N_b}$ is the nominal significance, but the ratio of cumulative distributions can improve it. The improvement factor in the significance from a cut on the likelihood is then

$$\Delta \mathcal{S} = \frac{1 - \Sigma_s(\mathcal{L})}{\sqrt{1 - \Sigma_b(\mathcal{L})}} = \frac{1 - \text{ROC}(x)}{\sqrt{1 - x}} \,, \tag{4.33}$$

where on the right, we have related the cumulative distributions of the likelihood ratio to the ROC curve. For quark versus gluon discrimination, this significance improvement factor is then

$$\Delta S = \frac{1 - x^{C_A/C_F}}{\sqrt{1 - x}} \,. \tag{4.34}$$

For $C_A/C_F = 9/4$ in QCD, this is maximized at $x \approx 0.41$, at which $\Delta S \approx 1.13$. So, there is gluon discovery improvement power of only about 13% over the nominal expectation.

4.3 Incorporation Higher Order Effects

This analysis is but the starting point for quark versus gluon discrimination. Essentially, no matter what you do or measure on jets, discrimination performance is bounded from below by this simple Casimir scaling result. By contrast, if you find that your quark versus gluon discriminant performs much worse than this, there is definitely something else going on! Nevertheless, this discrimination power isn't that great, and suggests that we should look for ways to improve it. In this section, we will discuss a couple of physics points that arise at higher orders, and in later sections, will revisit the Lund plane and establish the optimal discriminant given this emission phase space.

4.3.1 Resolving Multiple Emissions

The first thing we could consider as an improvement to the discrimination power of the angularities exclusively is to measure additional observables that are correspondingly sensitive to additional emissions on the Lund plane. The angularities are sensitive to radiation about a single hard
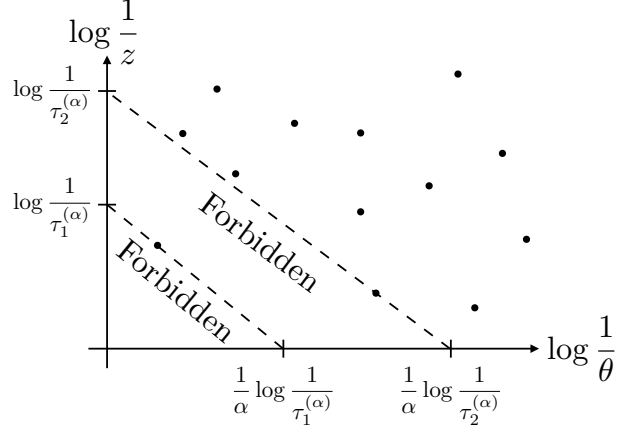
Figure 14: Representation on the Lund plane of the measurement of 1- and 2-subjettinesses, $\tau_1^{(\alpha)}$ and $\tau_2^{(\alpha)}$.

particle, or the jet axis, and in the soft and collinear limit, a single emission dominates. This suggests that if we want sensitivity to additional emissions in the jet, then we should consider radiation about additional axes within the jet. The observable that accomplishes this is called N-subjettiness [358, 359] (see also Refs. [360, 361, 362]), in which one places N axes in the jet along the directions of the N most dominant energy flows and then calculates the effect of additional radiation about those axes:

$$\tau_N^{(\alpha)} = \sum_{i \in J} z_i \, \min \{\theta_{i1}^\alpha, \theta_{i2}^\alpha, \dots, \theta_{iN}^\alpha\} \,, \tag{4.35}$$

where again $\alpha > 0$ for IRC safety. Here, θ_{iK} is the angle between particle i and axis K in the jet.

Let's consider the effect on quark versus gluon discrimination on jets on which we measure both the angularity $\tau_\alpha = \tau_1^{(\alpha)}$, which is equivalently 1-subjettiness, and 2-subjettiness, $\tau_2^{(\alpha)}$. What is especially convenient in the soft and collinear limit of the Lund plane is that the N-subjettiness axes always lie exactly on the first N particles ordered in their contribution to the angularity τ_α. This makes calculation of the joint distribution of $\tau_1^{(\alpha)}$ and $\tau_2^{(\alpha)}$ very simple. In Fig. 14, we have illustrated the Lund plane in which both $\tau_1^{(\alpha)}$ and $\tau_2^{(\alpha)}$ are measured.[20] This figure illustrates everything we need to know to directly write down the joint probability distribution. First, the total forbidden emission area is set by the value of $\tau_2^{(\alpha)}$, and so the Sudakov factor has no explicit dependence on $\tau_1^{(\alpha)}$. Next, there are two explicit emissions that set $\tau_1^{(\alpha)}$ and $\tau_2^{(\alpha)}$, and, as emissions are uniform on the Lund plane, the functional form of the differential probability of these emissions is identical. Therefore, we can immediately write down the joint probability distribution as

$$p_q(\tau_1^{(\alpha)}, \tau_2^{(\alpha)}) = \left(\frac{2\alpha_s C_F}{\pi \alpha}\right)^2 \frac{\log \tau_1^{(\alpha)}}{\tau_1^{(\alpha)}} \frac{\log \tau_2^{(\alpha)}}{\tau_2^{(\alpha)}} \exp\left[-\frac{\alpha_s}{\pi} \frac{C_F}{\alpha} \log^2 \tau_2^{(\alpha)}\right] \,. \tag{4.36}$$

---

[20]We are just considering emissions off of the leading, hard particle. A complete calculation for 2-subjettiness would also need to include emission off of the particle that set the value of 1-subjettiness. However, our simpler calculation is sufficient for the conclusions we wish to draw here.

As a consistency check of this distribution, we can integrate over $0 < \tau_2^{(\alpha)} < \tau_1^{(\alpha)}$ and verify that the correct distribution of $\tau_1^{(\alpha)}$ is produced:

$$\int_0^{\tau_1^{(\alpha)}} d\tau_2^{(\alpha)} \, p_q(\tau_1^{(\alpha)}, \tau_2^{(\alpha)}) = -\frac{2\alpha_s C_F}{\pi\alpha} \frac{\log \tau_1^{(\alpha)}}{\tau_1^{(\alpha)}} \exp\left[-\frac{\alpha_s}{\pi} \frac{C_F}{\alpha} \log^2 \tau_1^{(\alpha)}\right], \tag{4.37}$$

which agrees with what we calculated earlier for the angularities.

As earlier, there's no need for another calculation with gluon jets, we just need to replace $C_F \to C_A$:

$$p_g(\tau_1^{(\alpha)}, \tau_2^{(\alpha)}) = \left(\frac{2\alpha_s C_A}{\pi\alpha}\right)^2 \frac{\log \tau_1^{(\alpha)}}{\tau_1^{(\alpha)}} \frac{\log \tau_2^{(\alpha)}}{\tau_2^{(\alpha)}} \exp\left[-\frac{\alpha_s}{\pi} \frac{C_A}{\alpha} \log^2 \tau_2^{(\alpha)}\right]. \tag{4.38}$$

Correspondingly, the likelihood ratio is

$$\mathcal{L} = \frac{p_g(\tau_1^{(\alpha)}, \tau_2^{(\alpha)})}{p_q(\tau_1^{(\alpha)}, \tau_2^{(\alpha)})} = \left(\frac{C_A}{C_F}\right)^2 \exp\left[-\frac{\alpha_s}{\pi} \frac{C_A - C_F}{\alpha} \log^2 \tau_2^{(\alpha)}\right]. \tag{4.39}$$

We immediately see two properties of this more differential likelihood ratio. First, the likelihood is monotonic in $\tau_2^{(\alpha)}$, so $\tau_2^{(\alpha)}$ itself is the optimal discriminant. Next, note that the reducibility factor for the gluon, the maximum value of the likelihood, has increased with respect to just measuring an angularity. Now, the maximal purity of a gluon sample is restricting to the region where $\tau_2^{(\alpha)} \to \tau_1^{(\alpha)} \to 1$, and where gluon and quark jets are found in the ratio $C_A^2 : C_F^2$. This latter point is intriguing: by resolving two emissions, the gluon reducibility factor is squared; could resolving three emissions result in the gluon reducibility factor cubed, from just measuring an angularity? If so, why don't we just measure all possible emissions on the Lund plane? We will return to this point in a couple of lectures, but for more details about following this N-subjettiness trail, see Refs. [103, 363].

Calculating the ROC curve isn't as simple in this case as it was for the angularities, but we can nevertheless calculate the distribution of the optimal discriminant, $\tau_2^{(\alpha)}$. We just need to marginalize over $\tau_1^{(\alpha)}$ to produce

$$p_q(\tau_2^{(\alpha)}) = \int_{\tau_2^{(\alpha)}}^1 d\tau_1^{(\alpha)} \, p_q(\tau_1^{(\alpha)}, \tau_2^{(\alpha)}) = -2\left(\frac{\alpha_s C_F}{\pi\alpha}\right)^2 \frac{\log^3 \tau_2^{(\alpha)}}{\tau_2^{(\alpha)}} \exp\left[-\frac{\alpha_s}{\pi} \frac{C_F}{\alpha} \log^2 \tau_2^{(\alpha)}\right]. \tag{4.40}$$

As always, the gluon distribution immediately follows. While its functional form is not enlightening, we can calculate the ROC curve in this case, which we have plotted in Fig. 15. This clearly illustrates that by resolving two emissions on the Lund plane, discrimination power is indeed improved throughout all of phase space. To quantify this, we can calculate the AUC with two resolved emissions, and the result is simple

$$\text{AUC} = \int_0^1 d\tau_2^{(\alpha)} \, p_q(\tau_2^{(\alpha)}) \, \Sigma_g(\tau_2^{(\alpha)}) = \frac{C_F^2(3C_A + C_F)}{(C_A + C_F)^3} \approx 0.23. \tag{4.41}$$

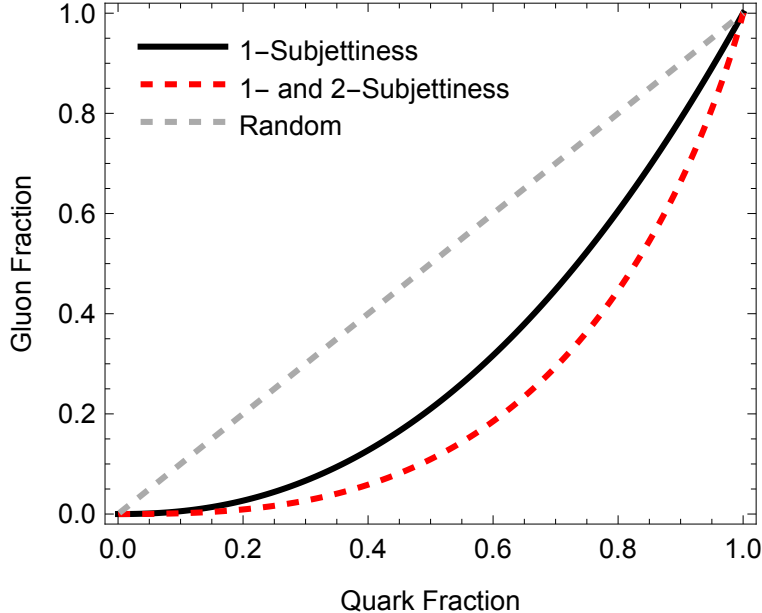Recall that this is smaller by about 0.08 than the AUC when just a single angularity is measured.

Figure 15: Comparison of the ROC curves for just measuring one angularity (black) to measuring 1- and 2-subjettiness, which resolves two emissions in the jet. We have also plotted the random AUC (gray), corresponding to the diagonal line.

4.3.2 The Importance of Being Recoil Free

Though we discussed recombination schemes in our aside on jet algorithms, we have been cavalier about the specific jet axis about which we measure the angularities τ_α in this section. At the double-logarithmic level, this ignorance and imprecision is fine, but at higher orders, the specific jet axis about which angularities are measured is important. Further, at the double logarithmic level, we saw no dependence of discrimination power on the specific value of the angular exponent α, but at higher orders this will be relevant.

Let's see how the jet axis and the angular exponent of the angularities τ_α come into play. We will just consider two axes as exemplar of the full range of possibilities: the E-scheme axis (the direction of mean flow of momentum) and the WTA axis, along the direction of the harder particle. Further, we will limit our analysis here to first non-trivial order, when the jets have but two particles in them. With this restriction, we will be able to use our master formula for jet observables to calculate appropriate distributions.

The first thing we need to do, however, is to determine the functional form of the angularities measured about these two different axes. In general, the angularities on a jet with two particles can be expressed as

$$\tau_\alpha = z\theta_1^\alpha + (1 - z)\theta_2^\alpha \,, \tag{4.42}$$

where z $(1 - z)$ is the energy fraction of particle 1 (2), and θ_1 (θ_2) is its angle from the appropriate axis. For the net momentum or E-scheme axis, the angle to the axis is set by the angle between

particles θ times the energy fraction of the other particle. Specifically,

$$\theta_1 = (1-z)\theta\,, \qquad\qquad\qquad \theta_2 = z\theta\,. \qquad\qquad (4.43)$$

With this, we can then evaluate the angularity about the E-scheme axis to be

$$\tau_\alpha^{(E\text{-scheme})} = z(1-z)^\alpha\theta^\alpha + z^\alpha(1-z)\theta^\alpha \xrightarrow{z\to 0} \begin{cases} z\theta^\alpha\,, & \alpha > 1\,, \\ 2z\theta\,, & \alpha = 1\,, \\ (z\theta)^\alpha\,, & \alpha < 1\,. \end{cases} \qquad (4.44)$$

At right, we have taken the soft, $z \to 0$, limit of the expression, which demonstrates some rather interesting behavior as a function of exponent α.

In the soft limit, for $\alpha > 1$, the angularity takes the "expected" form, $\tau_\alpha = z\theta^\alpha$, where the dynamics of the soft particle about the hard particle dominate. Something strange happens when $\alpha = 1$, and for $\alpha < 1$, the term that dominates in the soft limit is the recoil of the hard particle from the jet axis due to momentum conservation with the soft particle. This is then no longer directly sensitive to the dynamics of the soft particle, and is somewhat rather random and arbitrary, because additional radiation can push the hard particle around further. Additionally, note that the soft limit of the angularity for $\alpha < 1$ is monotonically related to the angularity with $\alpha = 1$. For discrimination performance, this means that angularities τ_α with $\alpha < 1$ will have identical discrimination power to the angularity with $\alpha = 1$. If there is sensitivity to α in discrimination performance, then that performance saturates at and below $\alpha = 1$.

By contrast, let's consider the angularity measure about the WTA axis. This expression is exceptionally simple because the WTA axis lies on the harder particle, so only the softer particle contributes:

$$\tau_\alpha^{(\text{WTA})} = \min[z, 1-z]\,\theta^\alpha \xrightarrow{z\to 0} z\theta^\alpha\,. \qquad\qquad (4.45)$$

Again, at right, we have taken the soft $z \to 0$ limit, but unlike for the E-scheme axis, the functional form of the WTA axis expression is the same for all $\alpha > 0$. Thus, if there is dependence on α in the discrimination power, we expect that there is no saturation at $\alpha = 1$ for angularities measured about the WTA axis.

Let's now calculate the distribution of the angularities on quark and gluon jets with two particles. From this analysis, we see that we only need to calculate the distribution with functional form $\tau_\alpha = z\theta^\alpha$, and further, we will demand that the particles are within an angle R. For quark jets, we have

$$\frac{d\sigma_q}{d\tau_\alpha} = \frac{\alpha_s C_F}{\pi} \int_0^R \frac{d\theta}{\theta} \int_0^1 dz\, \frac{1 + (1-z)^2}{z} \delta(\tau_\alpha - z\theta^\alpha) \qquad (4.46)$$

$$= -\frac{2\alpha_s C_F}{\pi\alpha} \frac{1}{\tau_\alpha}\left(\log\frac{\tau_\alpha}{R^\alpha} + \frac{3}{4}\right)\,,$$

where we have only kept the leading terms in the $\tau_\alpha \to 0$ limit. The corresponding calculation for gluon jets is

$$\frac{d\sigma_g}{d\tau_\alpha} = \frac{\alpha_s}{\pi} \int_0^R \frac{d\theta}{\theta} \int_0^{1/2} dz\, \left[2C_A\left(\frac{z}{1-z} + \frac{1-z}{z} + z(1-z)\right) + 2n_f T_R\left(z^2 + (1-z)^2\right)\right] \delta(\tau_\alpha - z\theta^\alpha)$$

$$= -\frac{2\alpha_s}{\pi\alpha}\frac{1}{\tau_\alpha}\left(C_A\log\frac{\tau_\alpha}{R^\alpha} + \frac{11}{12}C_A - \frac{n_f T_R}{3}\right). \tag{4.47}$$

Here, we have introduced n_f as the number of active quarks to which the initial gluon can split (the number of quarks whose mass is less than half the initial energy of the gluon).

Both of these expressions will have to be dressed up with their appropriate Sudakov factor to account to arbitrary additional gluon emission, but we can understand where α dependence in discrimination power arises without any more detailed calculations. Note that in both distributions, the α dependence is an overall factor; however, with the full splitting functions, there is a logarithmic term and a non-logarithmic term, and depending on the value of α, these can play off one another in interesting ways. To see how this works, let's stare at the Sudakov factor (for the quark), and make some estimates. The quark's Sudakov factor is

$$\Sigma_q(\tau_\alpha) = e^{-\frac{\alpha_s}{\pi}\frac{C_F}{\alpha}\log^2\tau_\alpha}. \tag{4.48}$$

As an exponential, most of its support lies where its argument is order-1, and so we can estimate that the Sudakov restricts the distribution to where

$$\frac{\alpha_s}{\pi}\frac{C_F}{\alpha}\log^2\tau_\alpha \sim 1. \tag{4.49}$$

Just focusing on the relationship between α and τ_α, this implies that

$$\log\tau_\alpha \sim -\sqrt{\alpha}, \tag{4.50}$$

at the value of α where the bulk of the distribution lies.

Using this scaling, let's plug it into the leading order distributions we had just calculated. We can ignore overall factors common to both quarks and gluons and just focus on the terms in parentheses. Note that we have the scaling

$$\frac{d\sigma_q}{d\tau_\alpha} \sim C_F\log\frac{\tau_\alpha}{R^\alpha} + \frac{3}{4}C_F \sim -\sqrt{\alpha} + \frac{3}{4}C_F, \tag{4.51}$$

$$\frac{d\sigma_g}{d\tau_\alpha} \sim C_A\log\frac{\tau_\alpha}{R^\alpha} + \frac{11}{12}C_A - \frac{n_f T_R}{3} \sim -\sqrt{\alpha} + \frac{11}{12}C_A - \frac{n_f T_R}{3}.$$

As $\alpha \to \infty$, the logarithm term dominates the distribution and therefore the discrimination power of the angularities reduces to the Casimir scaling that we established earlier. However, as α decreases, the importance of the logarithmic term decreases, and the new, constant, terms become more important. The values of the constant terms differ significantly in QCD, namely

$$\frac{3}{4}C_F = 1, \tag{4.52}$$

$$\frac{11}{12}C_A - \frac{n_f T_R}{3} = \frac{23}{12}, \tag{4.53}$$

where we have set $n_f = 5$, and so as their importance increases, so too does discrimination power. Therefore, as mentioned earlier, we expect that the WTA angularities are better quark versus gluon discriminants at small α than the E-scheme angularities, whose performance saturates around $\alpha = 1$.
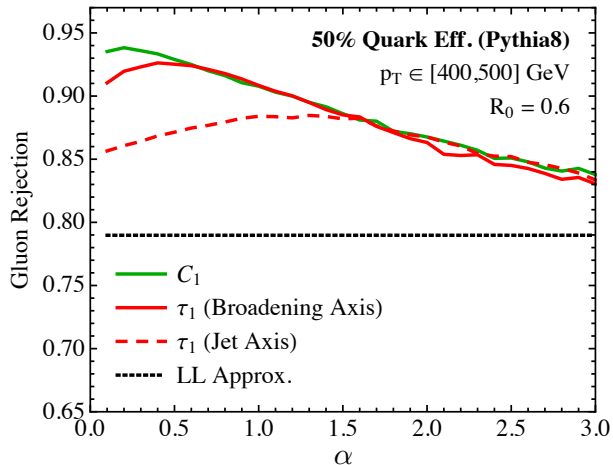
Figure 16: Plot of the gluon jet rejection rate (1−gluon efficiency) at 50% quark jet efficiency, as a function of angular exponent in the angularities, α, in simulated events at the LHC. The dashed black line corresponds to the Casimir scaling expectation, while the dashed red line is the rejection with angularities measured about the E-scheme jet axis, and the solid lines are the rejection rate for angularities measured about a recoil-free axis. Better discrimination power is higher on this plot. Slightly modified version of a plot from Ref. [298].

This observation is illustrated in simulated data in Fig. 16, reproduced from Ref. [298]. This plot shows the gluon jet rejection rate (1−gluon jet efficiency) at 50% quark jet efficiency, as a function of the angular exponent α of the angularity. The black dashed line is the Casimir scaling result at double logarithmic accuracy, which is independent of α. The color curves show the rejection rate for angularities measured about different axes: the E-scheme axis angularities are dashed red, and recoil-free axis angularities are solid. The WTA axis was not yet invented at the time of this reference, but the solid red curve (the "broadening axis") is in practice nearly identical to the WTA axis, and is indeed the same for a jet with two particles. As α decreases, we indeed see better discrimination power (higher rejection rate), but the E-scheme angularities stop improving below about $\alpha \sim 1.5$, close to the $\alpha = 1$ point we had predicted. By contrast, the recoil-free axis angularities continue to improve discrimination power down to very small values of α, where other physical effects start to dominate.[21]

Exercises

1. Consider the measurement of two angularities, τ_α and τ_β, with, say $\alpha > \beta$.

    (a) Calculate the Sudakov form factor for two angularities measured on quark jets, the joint probability distribution $p_q(\tau_\alpha, \tau_\beta)$. Further, ensure that the joint probability distribution

---

[21]These effects are likely dominantly non-perturbative in origin. Recall that in the Sudakov form factor, the coupling α_s and the angular exponent α appear in the ratio α_s/α. A perturbative analysis assumes that $\alpha_s/\alpha \ll 1$, so that truncation of the expansion is justified; however, at sufficiently small α, we will have $\alpha_s/\alpha \sim 1$, and perturbation theory breaks down.

marginalizes to the correct single probability distributions. That is,

$$\int_{\tau_{\text{lo}}}^{\tau_{\text{hi}}} d\tau_\beta \, p_q(\tau_\alpha, \tau_\beta) = p_q(\tau_\alpha) \,, \tag{4.54}$$

for particular bounds $\tau_{\text{lo}} < \tau_\beta < \tau_{\text{hi}}$ (that you should determine). For a hint to this problem, see Ref. [364].

(b) Determine the likelihood ratio for quark versus gluon discrimination from this joint probability distribution,

$$\mathcal{L} = \frac{p_g(\tau_\alpha, \tau_\beta)}{p_q(\tau_\alpha, \tau_\beta)} \,. \tag{4.55}$$

What is the gluon reducibility factor with this likelihood ratio?

(c) Can you determine the distribution of the likelihood ratio on quark and gluon jets? What is the ROC curve and how does it compare to the case when only a single angularity is measured?

2. (This is an extension of Exercise 9.3 in Ref. [53].) The ALEPH experiment at the Large Electron-Positron Collider (LEP) measured the number of jets produced in $e^+e^- \to$ hadrons collision events. The experiment counted n jets, if, for every pair i, j of jets the following inequality is satisfied:

$$2 \min[E_i^2, E_j^2](1 - \cos\theta_{ij}) > y_{\text{cut}} E_{\text{cm}}^2 \,, \tag{4.56}$$

for $y_{\text{cut}} < 1$, E_i is the energy of jet i, θ_{ij} is the angle between jets i and j and E_{cm} is the center-of-mass collision energy. In the soft and collinear limits, determine the probability p_n for observing n jets, as a function of y_{cut}.

Note that the minimum number of jets is 2 ($e^+e^- \to q\bar{q}$) and gluons can be emitted from either the quark or the anti-quark. Compare your result to figure 7 of Ref. [365]. What value of α_s fits the data the best? This fit is imperfect because we're omitting a lot of important physics, but it will be qualitatively close.

3. Consider the number of jets as defined by the procedure introduced in Exercise 2 above. Consider that procedure measured on $e^+e^- \to q\bar{q} + X$ and $e^+e^- \to gg + X$ events, where X is any other hadronic activity. Using the discrete probability distribution p_n for the quark and gluon final states, determine the likelihood ratio, ROC curve and AUC for this number of jets observable for discrimination of the quark from gluon final states, as a function of the parameter y_{cut}. Does the AUC for this observable ever correspond to better discrimination to that for τ_α as derived in this section?

5 Second Example: $H \to b\bar{b}$ Identification and IRC Safe Binary Discrimination

By the mid-2000s, just a year or two before the LHC was scheduled to turn on, the ATLAS and CMS experiments had clearly identified the path they would take to achieve the central goal of the LHC program: discovery of the Higgs boson. Much of that story is well known because, of course, the Higgs was discovered in 2012 [356, 357] through observation in the "golden channels"

of $H \to \gamma\gamma$ and $H \to 4$ leptons, which have clean experimental signatures and very small Standard Model backgrounds.[22] However, Higgs bosons only decay to photons or four leptons (through off-shell Z bosons) about 0.5% of the time combined, so just by discovering the Higgs in these modes, one can clearly not claim victory in knowing that what you saw was actually *the* Higgs. Most of the time, the vast majority of the time, nearly 90% of all decays of the Higgs produce hadrons in relatively low-energy jets (a scale around half the Higgs mass), but hadrons at these low energies are produced in huge quantities at the LHC.

Initial analyses were done by the experiments to quantify the status of searching for the Higgs in these hadronic decay modes, but the conclusion was that the backgrounds are simply too large. In particular, the dominant decay mode of the Higgs, $H \to b\bar{b}$ which occurs about 50% of the time itself, was not considered viable to observe at the LHC in the ATLAS and CMS Technical Design Reports [367, 368]. Then, again in February 2008 (perhaps the historically most important month for jet substructure!),[23] a group composed of theorists and experimentalists introduced a new technique that rendered the bottom decay mode of the Higgs observable [370], and subsequently kick-started the modern jet substructure endeavor.[24] The key observations of Ref. [370] were two-fold: first, Standard Model backgrounds decrease much more rapidly than Higgs production at higher energies, and second, that the inside of a "Higgs jet" should contain two hard prongs (the two bottom quarks), but multiple prong substructure was rare in QCD (essentially because of approximate scale invariance). Exploiting these features enabled the authors to show that discovery significances of 5 or more were possible. It took another decade, but direct evidence for $H \to b\bar{b}$ decays has now been observed [374, 375].

The $H \to b\bar{b}$ decay has remained an interesting and widely studied benchmark case, especially for machine learning discrimination studies (see, e.g., Refs. [376, 377, 378, 379, 380, 381, 382, 383, 384, 385, 386]). The dominant background to this decay at high energies relevant for jet substructure is the splitting of an off-shell gluon to bottom quarks, $g \to b\bar{b}$. No other hadronic backgrounds really need to be considered because bottom quark tagging in experiment is now nearly perfect [387, 388]. So, we can formulate the identification of $H \to b\bar{b}$ decays at high energies as a binary discrimination problem. However, unlike for our quark versus gluon discrimination in the previous section, the jets of interest now are definitely not scale invariant because the Higgs is massive. Instead, what we will do is work within fixed-order perturbation theory, identifying signal and background distributions on phase space, their likelihood ratio, and the corresponding discrimination power. Because of the explicit scale of the Higgs mass and the requirement of bottom quarks in the problem, leading-order distributions calculated in fixed-order perturbation theory will actually be true probabilities because there is no degeneracy of states that needs to be accounted for.

---

[22]The story of the discovery of the Higgs is documented in the fantastic film "Particle Fever", starring and produced by physicist David E. Kaplan, himself an author on a very early top tagging algorithm [366]. I happened to be at CERN on 4 July, 2012, but had caught a cold on the trans-Atlantic flight and was therefore unable to stay up all night saving a precious early spot in line for entrance into the auditorium. I was thus resigned to watching the discovery announcement from my computer in my hostel room.

[23]In addition to the anti-k_T algorithm and the BDRS mass drop-filtering algorithm for $H \to b\bar{b}$ identification, February 2008 also saw the posting of a method for removing pile-up radiation from jets [369], and then in early March 2008, a re-evaluation of energy correlators from a strong coupling/string theory perspective [339].

[24]See also Refs. [371, 372, 373] for historical foundations. As an undergraduate, I worked on a project starting in fall 2016 on a closely related problem, but it wasn't published and I then moved on to graduate school. See my abstract for a presentation at the 2007 University of Washington Undergraduate Research Symposium on page 150 of https://uw-s3-cdn.s3.us-west-2.amazonaws.com/wp-content/uploads/sites/73/2014/06/24075140/2007SympProceedings.pdf

First, however, we will formulate the binary discrimination problem in general in fixed-order perturbation theory, without assuming a particular problem or goal until later. Thus, $H \to b\bar{b}$ versus $g \to b\bar{b}$ discrimination will be the exemplar of this formalism, but can be applied more broadly. As another interesting case, significant work has also been devoted to developing techniques for discrimination of hadronic top quark decays from jets initiated by light QCD partons. In 2019, a review and comparison of a large number of machine learning approaches to this problem were studied and summarized in Ref. [389]. One conclusion of this study is that top quark discrimination and identification algorithms are now extremely powerful, but another conclusion is that all of these machine learning approaches (which at least naively differ significantly in their implementation, internal architecture, etc.) have very similar performance. Further, combining different algorithms seems to not improve performance very much, and so suggests that there is a maximal upper bound to possible discrimination. Machine learning cannot answer this question, because the models are all trained on simulated data, but from a principled, order-by-order analysis of the problem, we can hope to do so.[25] Top quark discrimination is much more challenging than that for Higgs bosons (at leading order, top quarks decay to three final state particles), but this framework can be extended to that case.

While we will be able to understand $H \to b\bar{b}$ decays versus $g \to b\bar{b}$ splittings in detail, there are other, related searches and discrimination problems for which the goal or signal is less well-defined. In searches for new physics or new particles, the mass of the particle is unknown, and so one needs a more general framework or reduced set of assumptions to construct the likelihood ratio. In this lecture, we will also review and apply power counting methods to optimal observable construction, given a parametric and hierarchical picture of what signal and background are expected to be, with no strict assumption on a particular value of the mass. Through this, we will also see our first glimpses beyond the cosy confines of perturbation theory, and demonstrate that seemingly innocuous questions have shocking answers. But first, to binary discrimination in fixed-order perturbation theory.

5.1 Fixed-Order Expansion of the Likelihood Ratio

We will study the general problem of binary discrimination in fixed-order perturbation theory, building up a description of the likelihood ratio, ROC curve, etc., in powers of the coupling α_s. The analysis presented here and in the specific example of $H \to b\bar{b}$ identification later closely follows that from Ref. [390], where more details can be found, including a complete calculation at next-to-leading order. Our starting point will be understanding the signal and background distributions on phase space Φ. We can generically write the signal s and background b distributions as a series in α_s as

$$\tilde{p}_s(\Phi) = p_s^{(0)}(\Phi^{(0)}) + \frac{\alpha_s}{2\pi} p_s^{(1)}(\Phi^{(1)}) + \mathcal{O}(\alpha_s^2) \,, \tag{5.1}$$

$$\tilde{p}_b(\Phi) = p_b^{(0)}(\Phi^{(0)}) + \frac{\alpha_s}{2\pi} p_b^{(1)}(\Phi^{(1)}) + \mathcal{O}(\alpha_s^2) \,, \tag{5.2}$$

just working through next-to-leading order explicitly. Here, the over-tilde represents that (so far) these are unnormalized distributions, which we will work to fix up. The superscript (n) denotes the order in α_s of that contribution, with (0) being leading order, (1) next-to-leading order, etc. If

---

[25]I thank Gregor Kasieczka for inspiring my interest in this problem.

there is any hope that these are probability distributions, the leading order distributions must be integrable themselves. For simplicity, we assume that they are also already normalized, where

$$1 = \int d\Phi^{(0)} \, p(\Phi^{(0)}) \,. \tag{5.3}$$

This will make future expressions significantly more compact.

We will make the additional assumption that the signal and background final states are indistinguishable. That is, at each order in perturbation theory and on each corresponding phase space $\Phi^{(n)}$, signal and background events have the same particle content. Thus, as quantum mechanical processes, we should coherently sum the signal and background states together, and the take their combined absolute square. This then suggests a related discrimination problem of unambiguously observing the quantum interference of the signal and background states. This problem can again be formulated as a binary discrimination problem, distinguishing the quantum pure state density matrix of signal from the classically mixed density matrix of background [391]. While several avenues for quantum interference at colliders have been proposed, from interference of Standard Model processes with new physics to interference between intermediate on-shell gluon spin states, e.g., Refs. [392, 393, 394, 395, 396, 397, 398, 399, 400, 343, 401, 402, 403], we won't study this fascinating direction further here.

While the leading-order distribution is normalized, the full distribution $\tilde{p}(\Phi)$ is not, so let's work to calculate the normalization factor $\mathcal{N}$ that accomplishes this. We require that

$$\begin{aligned}
1 &= \mathcal{N} \int d\Phi \, \tilde{p}(\Phi) \tag{5.4} \\
&= \left(1 + \frac{\alpha_s}{2\pi} \mathcal{N}^{(1)} + \cdots\right) \left(1 + \frac{\alpha_s}{2\pi} \int d\Phi^{(1)} p^{(1)}(\Phi^{(1)}) + \cdots\right) \\
&= 1 + \frac{\alpha_s}{2\pi} \left(\int d\Phi^{(1)} \, p^{(1)}(\Phi^{(1)}) + \mathcal{N}^{(1)}\right) + \cdots.
\end{aligned}$$

On the second line, we have introduced the next-to-leading order normalization factor $\mathcal{N}^{(1)}$ and ellipses hide terms at higher orders in α_s. Through next-to-leading order, then, the normalization factor has the expansion

$$\mathcal{N} = 1 - \frac{\alpha_s}{2\pi} \int d\Phi^{(1)} \, p^{(1)}(\Phi^{(1)}) + \cdots. \tag{5.5}$$

This can then be used to construct unit normalized distributions on phase space (no over-tildes) as

$$\begin{aligned}
p(\Phi) &= \mathcal{N} \tilde{p}(\Phi) \tag{5.6} \\
&= p^{(0)}(\Phi^{(0)}) + \frac{\alpha_s}{2\pi} p^{(1)}(\Phi^{(1)}) - \frac{\alpha_s}{2\pi} p^{(0)}(\Phi^{(0)}) \int d\Phi^{(1)} \, p^{(1)}(\Phi^{(1)}) + \cdots.
\end{aligned}$$

With normalized probability distributions, we can then evaluate the likelihood ratio on phase space to establish the optimal observable possible, given that we only measure particle momenta. The form of the likelihood ratio we will consider here is the inverse of that we introduced with the Neyman-Pearson Lemma (signal distribution is in the denominator now), which we will find is convenient when applied to Higgs decays. As a function of phase space variables, the likelihood ratio is

$$\hat{\mathcal{L}}(\Phi) = \frac{p_b(\Phi)}{p_s(\Phi)} \tag{5.7}$$

$$
= \frac{p_b^{(0)}(\Phi^{(0)}) + \frac{\alpha_s}{2\pi} p_b^{(1)}(\Phi^{(1)}) - \frac{\alpha_s}{2\pi} p_b^{(0)}(\Phi^{(0)} \int d\Phi^{(1)} p_b^{(1)}(\Phi^{(1)})) + \cdots}{p_s^{(0)}(\Phi^{(0)}) + \frac{\alpha_s}{2\pi} p_s^{(1)}(\Phi^{(1)}) - \frac{\alpha_s}{2\pi} p_s^{(0)}(\Phi^{(0)}) \int d\Phi^{(1)} p_s^{(1)}(\Phi^{(1)}) + \cdots}
$$

$$
= \frac{p_b^{(0)}(\Phi^{(0)})}{p_s^{(0)}(\Phi^{(0)})} \left[1 + \frac{\alpha_s}{2\pi} \left(\int d\Phi^{(1)} \left(p_s^{(1)}(\Phi^{(1)}) - p_b^{(1)}(\Phi^{(1)}) \right) + \frac{p_b^{(1)}(\Phi^{(1)})}{p_b^{(0)}(\Phi^{(0)})} - \frac{p_s^{(1)}(\Phi^{(1)})}{p_s^{(0)}(\Phi^{(0)})} \right) + \cdots \right] ,
$$

expanding through next-to-leading order on the final line. We can now identify terms in the Taylor expansion of the likelihood as an observable itself:

$$
\hat{\mathcal{L}}(\Phi) = \hat{\mathcal{L}}^{(0)}(\Phi) + \frac{\alpha_s}{2\pi} \hat{\mathcal{L}}^{(1)}(\Phi) + \cdots , \tag{5.8}
$$

where

$$
\hat{\mathcal{L}}^{(0)}(\Phi) = \frac{p_b^{(0)}(\Phi^{(0)})}{p_s^{(0)}(\Phi^{(0)})} , \tag{5.9}
$$

$$
\hat{\mathcal{L}}^{(1)}(\Phi) = \frac{p_b^{(0)}(\Phi^{(0)})}{p_s^{(0)}(\Phi^{(0)})} \left[\int d\Phi^{(1)} \left(p_s^{(1)}(\Phi^{(1)}) - p_b^{(1)}(\Phi^{(1)}) \right) + \frac{p_b^{(1)}(\Phi^{(1)})}{p_b^{(0)}(\Phi^{(0)})} - \frac{p_s^{(1)}(\Phi^{(1)})}{p_s^{(0)}(\Phi^{(0)})} \right] , \tag{5.10}
$$

so the likelihood ratio as the optimal observable itself is modified order-by-order.

If this likelihood ratio is to be well-defined and IRC safe, we now see immediately some requirements it must satisfy. First and foremost, the leading order distribution ratio

$$
\hat{\mathcal{L}}^{(0)}(\Phi) = \frac{p_b^{(0)}(\Phi^{(0)})}{p_s^{(0)}(\Phi^{(0)})} , \tag{5.11}
$$

must be well-defined and finite on leading-order phase space $\Phi^{(0)}$ if this is to make any sense whatsoever. Starting at next-to-leading order, there are terms in the likelihood ratio that mix leading and next-to-leading order distributions which correspondingly live on different phase spaces, $\Phi^{(0)}$ and $\Phi^{(1)}$. If these are to be consistent and well-defined, we must map next-to-leading order phase space $\Phi^{(1)}$ onto leading order phase space $\Phi^{(0)}$ in an IRC safe way. We know how to do this: we can just use a sequential recombination jet algorithm to cluster the particles at next-to-leading order into the correct number of particles that live on leading order phase space. In the strict soft and collinear limits, IRC safety requires that this map is unambiguous and universal, but away from those limits, algorithm dependence will be introduced. We will not study effects of different algorithms here because they will require a complete and detailed calculation at next-to-leading order.

Regarding the functional form of the next-to-leading order contribution to the likelihood ratio in the square brackets, there are two distinct parts. One is the integral

$$
\int d\Phi^{(1)} \left(p_s^{(1)}(\Phi^{(1)}) - p_b^{(1)}(\Phi^{(1)}) \right) , \tag{5.12}
$$

which is simply a number that ensures proper normalization. As such, we won't consider it further here. The more interesting piece is the difference of the background to signal distribution ratios:

$$
\frac{p_b^{(1)}(\Phi^{(1)})}{p_b^{(0)}(\Phi^{(0)})} - \frac{p_s^{(1)}(\Phi^{(1)})}{p_s^{(0)}(\Phi^{(0)})} , \tag{5.13}
$$

which has explicit dependence on next-to-leading order phase space and so will be doing the heavy lifting of discrimination at this order.

Let's dig a bit deeper into these ratios of distributions on phase space. We can immediately write down explicit expressions for these ratios from our understanding of soft and collinear factorization from Sec. 3.2. From our results on soft and collinear factorization, we can express the ratio of next-to-leading to leading order distributions in the following way:

$$\frac{p^{(1)}(\Phi^{(1)})}{p^{(0)}(\Phi^{(0)})} = \sum_{\text{NLO partons } i,j} \frac{(4\pi)^2}{s_{ij}} P_{(ij)\to ij}(z) - \sum_{\text{LO partons } i,j} (4\pi)^2 \mathbf{T}_i \cdot \mathbf{T}_j \frac{s_{ij}}{s_{ik}s_{kj}} + \text{non-singular} \,.$$

(5.14)

There are a few things going on, so let's explain. The first sum encodes all hard, collinear emissions, as described by the invariant mass of particles i and j, s_{ij}, and corresponding splitting functions $P_{(ij)\to ij}(z)$. (We should also explicitly subtract the soft and collinear limit to avoid double counting with soft emissions, but we will see in a second that such terms cancel anyway.) The second sum encodes soft gluon emission k from every dipole i,j at leading order. Finally, furthest at right, terms in the ratio of distributions that are non-singular on phase space have not been explicitly written.

Let's now apply this factorization to understanding the difference between the background and signal ratio distributions. First, there is a significant cancelation we can take account of. Earlier, we had assumed that the signal and background states were indistinguishable and as such have the same particle content. As such, at next-to-leading order, the collinear emission contributions of the signal and background are identical, and therefore cancel in the difference. Collinear emissions, then, apparently do not modify the likelihood ratio, at least through next-to-leading order. Soft emissions are a different case because of sensitivity to the color flow along and connection amongst dipoles at leading order. At any rate, we can compactly represent the difference of the ratio as

$$\frac{p_b^{(1)}(\Phi^{(1)})}{p_b^{(0)}(\Phi^{(0)})} - \frac{p_s^{(1)}(\Phi^{(1)})}{p_s^{(0)}(\Phi^{(0)})} = -(4\pi)^2 \sum_{\text{LO partons } i,j} \left(\mathbf{T}_i^{(b)} \cdot \mathbf{T}_j^{(b)} - \mathbf{T}_i^{(s)} \cdot \mathbf{T}_j^{(s)} \right) \frac{s_{ij}}{s_{ik}s_{kj}} + \text{non-singular} \,.$$

(5.15)

Here, the superscripts on the color matrices represent whether they are from signal or background. We see then, completely generally, that the likelihood ratio at next-to-leading order is explicitly sensitive to differences in the flow of color from signal and background.

We could continue, and calculate the signal and background distributions of the likelihood ratio order-by-order in α_s, and then further combine those results into the ROC curve, but we won't do that here. Our observation in Eq. (5.15) that differences in color flow are the dominant contribution to the likelihood ratio at next-to-leading order will be sufficient to deeply understand $H \to b\bar{b}$ versus $g \to b\bar{b}$ discrimination.

5.2 $H \to b\bar{b}$ versus $g \to b\bar{b}$ and Sensitivity to Color Representation

For the problem at hand, we will make a number of simplifying assumptions so that our analysis is contained. The decay width the of Higgs boson is very small compared to its mass, $\Gamma_H \approx 3$ MeV compared to $m_H \approx 125$ GeV, so we can work in the narrow width approximation in which there is no quantum interference between signal and background. We will work in the highly-boosted limit,

in which the energy of the Higgs boson E is much larger than its mass, m_H. As such, the decay products of the Higgs boson will be highly collimated, and we can work in the collinear limit to describe the splitting of the gluon to bottom quarks. Additionally, in this highly boosted limit, the rest of the event is effectively an "infinite" angle away from the Higgs, and so can be considered as a single particle off of which the Higgs recoils. For an actual calculation, we will also need to use a jet algorithm and impose a jet radius about the Higgs and gluon, but we won't worry about that detail here. We will just assume that the jet radius is sufficiently large such that all decay products of the Higgs are comfortably contained within the jet.

Before doing any analysis or performing any calculation, you should always ask yourself what result you expect. Of course this can't necessarily be quantitative (for that is the point of a calculation!), but we can understand some limiting behavior and our results must respect that. Let's consider discrimination between $H \to b\bar{b}$ and $g \to b\bar{b}$ jets in the truly infinite boost limit, for $E/m_H \to \infty$. In this limit, the decay products of the Higgs boson are exactly collinear. Further, the Higgs boson is a color singlet, and so there is no net color observable in this Higgs. As such, there is no possible color dipole that can emit soft gluons a finite angle from the Higgs. So, a Higgs jet in the infinite boost limit just consists of a single, collimated core of energy. By contrast, the background gluon is a color octet, and so even in the infinite boost limit, its $b\bar{b}$ splitting products form a collinear color octet state. This then has a net color charge and therefore can emit soft gluons at finite angles. So, if you observe even a *single* soft gluon at any finite angle from the infinitely-boosted jet core, then that jet was necessarily initiated by a gluon. Thus, we expect that in the infinite boost limit, there is perfect discrimination of color-singlet Higgs bosons from color-octet gluons.

In an actual, realistic experiment like ATLAS or CMS at the LHC this thought experiment is a bit unrealistic because there are many other sources of radiation that might happen to deposit in the jet of interest, such as errant radiation off of other produced jets, underlying event radiation, initial state radiation, etc., so no perfect discrimination is ever expected. Nevertheless, we do expect that discrimination power improves as jet energy increases for the reasons explained above, which is something we can then test in simulated data with our machine learning algorithm, if so desired.

5.2.1 Discrimination at Leading Order

Let's now put all of this esoterica to use and actually calculate likelihoods and discrimination power for this problem. Working in the collinear limit, we know that relevant two-body phase space is two dimensional, and can be expressed as the invariant mass s of the particles and the energy fraction z of one of them. Because we are searching for a Higgs boson, we know its mass and so fix $s = m_H^2$ in this approximation, because if the mass of the jet is not m_H, there's no way that that jet is a Higgs. Then, there is but a single variable on leading order phase space to consider, the energy fraction z.

Because the Higgs boson decays on-shell and is a scalar, spin-0 particle, the probability distribution of the energy fraction of one of its decay products is flat in z:

$$p_H^{(0)}(z) = 1 \,, \tag{5.16}$$

which is normalized on $z \in [0,1]$. We are being slightly cavalier here, because there should be lower and upper bounds on the energy fraction as imposed by the constraint on the invariant mass. A truly 0 energy particle has no invariant mass with any other particle, however, in the high

energy limit, this restriction is suppressed by m_H^2/E^2, which we can safely ignore for simplicity. Correspondingly, for the background process $g \to b\bar{b}$, we can use the corresponding splitting function $P_{g \to q\bar{q}}(z)$, appropriately normalized on $z \in [0, 1]$:

$$p_g^{(0)}(z) = \frac{3}{2} \left[z^2 + (1 - z)^2 \right] . \tag{5.17}$$

Though the bottom quark is massive, $m_b \approx 4.2$ GeV, this is much less than the mass of the Higgs boson, $m_H \approx 125$ GeV, and so is safely ignorable for our purposes here.

Then, with these distributions, the leading-order likelihood ratio is

$$\hat{\mathcal{L}}^{(0)}(z) = \frac{p_g^{(0)}(z)}{p_H^{(0)}(z)} = \frac{3}{2} \left[z^2 + (1 - z)^2 \right] \in \left[\frac{3}{4}, \frac{3}{2} \right] , \tag{5.18}$$

and we clearly see why we put the signal Higgs in the denominator. This likelihood ratio has a very small dynamic range, just a factor of 2 from minimum to maximum, so purifying a sample of Higgs bosons only through measuring the energy fraction of one of the decay products is not possible. Nevertheless, for good exercise, let's continue on, and calculate the ROC curve. To do this, we need the cumulative distributions of the likelihood on signal and background:

$$\Sigma_H^{(0)}(\mathcal{L}) = \int_0^1 dz \, \Theta \left(\mathcal{L} - \frac{3}{2} \left[z^2 + (1 - z)^2 \right] \right) = \sqrt{\frac{4}{3} \mathcal{L} - 1} , \tag{5.19}$$

$$\Sigma_g^{(0)}(\mathcal{L}) = \int_0^1 dz \, \frac{3}{2} \left[z^2 + (1 - z)^2 \right] \Theta \left(\mathcal{L} - \frac{3}{2} \left[z^2 + (1 - z)^2 \right] \right) = \left(\frac{1}{2} + \frac{\mathcal{L}}{3} \right) \sqrt{\frac{4}{3} \mathcal{L} - 1} . \tag{5.20}$$

The leading order ROC curve is then

$$\text{ROC}^{(0)}(x) = \Sigma_g^{(0)} \left(\Sigma_H^{(0)^{-1}}(x) \right) = \frac{3}{4} x + \frac{x^3}{4} . \tag{5.21}$$

The AUC is correspondingly very large

$$\text{AUC}^{(0)} = \int_0^1 dx \, \text{ROC}^{(0)}(x) = \frac{7}{16} = 0.4375 , \tag{5.22}$$

close to completely random AUC = 0.5, and so at leading order there is very little discrimination power. What gives? We had argued that discrimination power should get arbitrarily good at high energies, but that argument required radiation off of the leading-order dipoles. Radiation only first arises at next-to-leading order, so let's go to higher order and see what we find!

5.2.2 Discrimination at Next-to-Leading Order and Beyond

At next-to-leading order, our signal jets now consist of $H \to b\bar{b}g$ final states and background jets consist of $g \to b\bar{b}g$ final states. As just mentioned, the additional radiated gluon will be doing the heavy lifting of discrimination at this order, which we will show shortly. Before that, however, I want to comment on something that we mentioned in the general expression for the likelihood ratio at next-to-leading order. We had found that we needed an IRC safe map or projection $\mathcal{P}$ from next-to-leading to leading order phase space: $\mathcal{P} : \Phi^{(1)} \to \Phi^{(0)}$. Not only must this map be

IRC safe, but further, the particle content at next-to-leading order must be consistently mapped to particles at leading order. For the case at hand, leading order jets are defined by having both a bottom b and anti-bottom $\bar{b}$ quark, and so our projection $\mathcal{P}$ must respect this. To do this requires enforcing that the b and $\bar{b}$ are resolved at next-to-leading order (i.e., with sufficient energy and a sufficiently large angle from each other). As such, the projection $\mathcal{P}$ always clusters the gluon with one of the bottoms, and never the bottoms together.

This prescription is fine and practical at next-to-leading order, but attempting its naive application at higher orders, one finds that flavor identification of the clustered subjets as bottom quarks is not IRC safe itself [300]. Soft quark emission at next-to-next-to-leading order can spoil simple flavor sums of clustered subjets, and so a more complicated procedure must be introduced to ensure calculability. Constructing and calculating with IR (and possibly C) safe flavor algorithms is a rather hot topic now [404, 405, 406, 407, 408], but we can ignore these subtleties at next-to-leading order, and just assume that whatever we do can be absorbed into a more principled approach at higher orders.

5.2.3 Likelihood Ratio in the Soft Limit

With those caveats out of the way, let's determine the likelihood ratio for $H \to b\bar{b}$ versus $g \to b\bar{b}$ at next-to-leading order. Let's focus on the component of the likelihood that is doing the heavy lifting for us, namely, the difference of ratios of distributions:

$$\frac{p_g^{(1)}(\Phi^{(1)})}{p_g^{(0)}(\Phi^{(0)})} - \frac{p_H^{(1)}(\Phi^{(1)})}{p_H^{(0)}(\Phi^{(0)})} = -(4\pi)^2 \sum_{\text{LO partons } i,j} \left(\mathbf{T}_i^{(g)} \cdot \mathbf{T}_j^{(g)} - \mathbf{T}_i^{(H)} \cdot \mathbf{T}_j^{(H)} \right) \frac{s_{ij}}{s_{ik}s_{kj}} + \text{non-singular} \,. \tag{5.23}$$

To evaluate this, we need to calculate the various color matrix products, but even before that, we need to establish what the sum over leading order particles is. At leading order there are, of course the bottom quarks, b and $\bar{b}$, but additionally, there is the "particle" that represents the rest of the event, off of which the Higgs or gluon recoils. We will call this particle $\bar{n}$ and, in the collinear limit, the direction of $\bar{n}$ is exactly opposite to that of the jet. Further, because the Higgs is a color singlet, the color of $\bar{n}$ in its event is 0, $\mathbf{T}_{\bar{n}}^{(H)} = 0$. Then, expanding out the sum over leading order partons, this contribution to the likelihood ratio is

$$\frac{p_g^{(1)}(\Phi^{(1)})}{p_g^{(0)}(\Phi^{(0)})} - \frac{p_H^{(1)}(\Phi^{(1)})}{p_H^{(0)}(\Phi^{(0)})} = -(4\pi)^2 \left(\mathbf{T}_b^{(g)} \cdot \mathbf{T}_{\bar{b}}^{(g)} - \mathbf{T}_b^{(H)} \cdot \mathbf{T}_{\bar{b}}^{(H)} \right) \frac{2s_{b\bar{b}}}{s_{bk}s_{k\bar{b}}} \tag{5.24}$$

$$- (4\pi)^2 \mathbf{T}_b^{(g)} \cdot \mathbf{T}_{\bar{n}}^{(g)} \frac{2s_{b\bar{n}}}{s_{bk}s_{k\bar{n}}} - (4\pi)^2 \mathbf{T}_{\bar{b}}^{(g)} \cdot \mathbf{T}_{\bar{n}}^{(g)} \frac{2s_{\bar{b}\bar{n}}}{s_{\bar{b}k}s_{k\bar{n}}} + \text{non-singular} \,.$$

Let's calculate some color matrix products! These cases are simple enough that we just need to use color conservation. Because the Higgs is a color singlet, the sum of its color matrices vanishes:

$$\left(\mathbf{T}_b^{(H)} + \mathbf{T}_{\bar{b}}^{(H)} \right)^2 = 0 = 2C_F + 2\mathbf{T}_b^{(H)} \cdot \mathbf{T}_{\bar{b}}^{(H)} \,, \tag{5.25}$$

because the color of a bottom quark is $C_F = \mathbf{T}_b^2$. The gluon is just a bit more complicated because there is an additional color direction. Nevertheless, color conservation requires

$$\left(\mathbf{T}_b^{(g)} + \mathbf{T}_{\bar{b}}^{(g)} + \mathbf{T}_{\bar{n}}^{(g)} \right)^2 = 0 = 2C_F + C_A + 2\mathbf{T}_b^{(g)} \cdot \mathbf{T}_{\bar{b}}^{(g)} + 2\mathbf{T}_b^{(g)} \cdot \mathbf{T}_{\bar{n}}^{(g)} + 2\mathbf{T}_{\bar{b}}^{(g)} \cdot \mathbf{T}_{\bar{n}}^{(g)} \,, \tag{5.26}$$
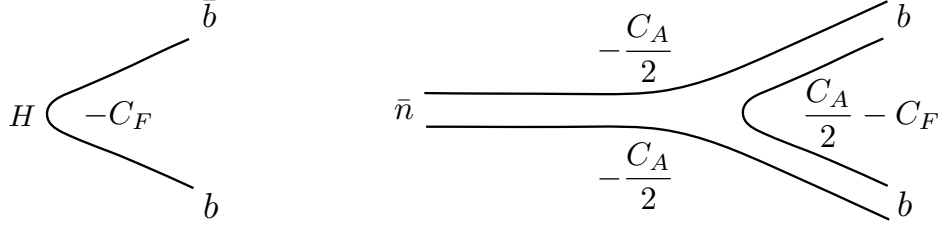
Figure 17: Illustrations of color flow between bottom quarks in Higgs decay (left) and in gluon splitting (right). Because the gluon is a color octet, there are dipole connections to the rest of the event, represented by the $\bar{n}$ direction. The lines indicate color matrix products between the particles at the endpoints. Figure adapted from Ref. [409].

where we have used that the color of the recoiling event must be $C_A = \mathbf{T}_{\bar{n}}^2$. Additionally, we have the constraints that the color products of the bottom and anti-bottom with $\bar{n}$ are equal $\mathbf{T}_b^{(g)} \cdot \mathbf{T}_{\bar{n}}^{(g)} = \mathbf{T}_{\bar{b}}^{(g)} \cdot \mathbf{T}_{\bar{n}}^{(g)}$, and that the color of the gluon is

$$\left(\mathbf{T}_b^{(g)} + \mathbf{T}_{\bar{b}}^{(g)}\right)^2 = C_A = 2C_F + 2\mathbf{T}_b^{(g)} \cdot \mathbf{T}_{\bar{b}}^{(g)} \,. \tag{5.27}$$

These results can then be simplified to evaluate all necessary color matrix products. These are summarized in Fig. 17, and then the expression for the difference of the ratio of distributions becomes

$$\frac{p_g^{(1)}(\Phi^{(1)})}{p_g^{(0)}(\Phi^{(0)})} - \frac{p_H^{(1)}(\Phi^{(1)})}{p_H^{(0)}(\Phi^{(0)})} = -(4\pi)^2 C_A \left(\frac{s_{b\bar{b}}}{s_{bk}s_{k\bar{b}}} - \frac{s_{b\bar{n}}}{s_{bk}s_{k\bar{n}}} - \frac{s_{\bar{b}\bar{n}}}{s_{\bar{b}k}s_{k\bar{n}}}\right) + \text{non-singular} \,. \tag{5.28}$$

Thus, the likelihood is explicitly sensitive to color of the gluon, C_A.

It is useful to further manipulate this expression to isolate the geometric dependence in the kinematic factor in parentheses. Recall that, in the collinear limit, the invariant mass of particles 1 and 2 is

$$s_{12} = z_1 z_2 E^2 \theta_{12}^2 \,, \tag{5.29}$$

where z_1 is the energy fraction of particle 1, E is the total jet energy, and θ_{12} is their pairwise angle. Additionally, we have to evaluate invariant masses with the rest of the event, e.g., $s_{b\bar{n}}$. By assumption in the high-boost limit, the b and $\bar{n}$ are back-to-back, and so this invariant mass is

$$s_{b\bar{n}} = 4E_b E_{\bar{n}} \,. \tag{5.30}$$

With these results, the contribution to the likelihood can be written as

$$\frac{p_g^{(1)}(\Phi^{(1)})}{p_g^{(0)}(\Phi^{(0)})} - \frac{p_H^{(1)}(\Phi^{(1)})}{p_H^{(0)}(\Phi^{(0)})} = \frac{(4\pi)^2 C_A}{z_k^2 E^2} \frac{\theta_{bk}^2 + \theta_{\bar{b}k}^2 - \theta_{b\bar{b}}^2}{\theta_{bk}^2 \theta_{\bar{b}k}^2} + \text{non-singular} \,. \tag{5.31}$$

In the collinear limit, we can then use the law of cosines to exchange $\theta_{b\bar{b}}^2$ for an opening angle ϕ in the triangle with vertices at the b, $\bar{b}$, and gluon k. We have

$$\theta_{b\bar{b}}^2 = \theta_{bk}^2 + \theta_{\bar{b}k}^2 - 2\theta_{bk}\theta_{\bar{b}k}\cos\phi \,, \tag{5.32}$$
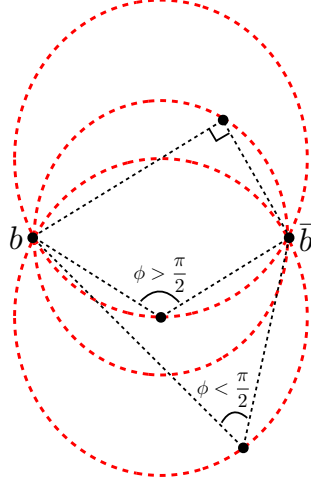
Figure 18: Regions selected by the structure of the next-to-leading order contribution to the likelihood ratio, with respect to the $b\bar{b}$ dipole. Between the $b\bar{b}$ dipole, radiation is suppressed ($\phi > \pi/2$ and so $\cos\phi < 0$) while outside the dipole, radiation is enhanced ($\phi < \pi/2$ and so $\cos\phi > 0$).

so that the difference can be nicely and compactly expressed as

$$\frac{p_g^{(1)}(\Phi^{(1)})}{p_g^{(0)}(\Phi^{(0)})} - \frac{p_H^{(1)}(\Phi^{(1)})}{p_H^{(0)}(\Phi^{(0)})} = \frac{2(4\pi)^2 C_A}{z_k^2 E^2} \frac{\cos\phi}{\theta_{bk}\theta_{\bar{b}k}} + \text{non-singular}. \tag{5.33}$$

5.2.4 Discrimination Power at Next-to-Leading Order

Let's now put this all together and write the expression for the likelihood ratio for $H \to b\bar{b}$ versus $g \to b\bar{b}$ discrimination through next-to-leading order:

$$\hat{\mathcal{L}}(\Phi) = \frac{3}{2}\left(z_b^2 + z_{\bar{b}}^2\right)\left(1 + (4\pi)^2\frac{\alpha_s}{2\pi}\frac{2C_A}{z_k^2 E^2}\frac{\cos\phi}{\theta_{bk}\theta_{\bar{b}k}} + \text{non-singular} + \mathcal{O}(\alpha_s^2)\right). \tag{5.34}$$

For the leading order likelihood, we have expressed the gluon splitting function $P_{g\to b\bar{b}}(z)$ as a function of the energy fractions of the bottom and anti-bottom, z_b and $z_{\bar{b}}$, respectively. At next-to-leading order, note that this is explicitly dimensionful, with the factor of squared jet energy in the denominator. However, this is canceled by the corresponding factor of squared energy from three-body collinear phase space of Eq. (3.26), where we also include the δ-function $\delta(s_{b\bar{b}k} - m_H^2)$ to constrain the jet mass appropriately. Further, when expressed in appropriate coordinates, three-body collinear phase space is flat in the squared angles of the emitted gluon to both bottom quarks and the squared energy of the soft gluon, $d\Phi^{(1)} \propto d\theta_{bk}^2\, d\theta_{\bar{b}k}^2\, dz_k\, z_k$ (see Exercise 1 of Sec. 3), and so there is no collinear divergence of the likelihood ratio at next-to-leading order. There is only a single soft divergence which is logarithmic, scaling as dz_k/z_k on phase space.

The dependence on the azimuthal angle ϕ in the term at next-to-leading order can be geometrically understood by drawing a picture of the $b\bar{b}$ dipole structure on the celestial sphere, as illustrated in Fig. 18. On this figure, we have drawn characteristic field value contours of a dipole,

which are also contours of the cosine of the opening angle ϕ of the sides of the triangle formed from the emitted gluon and the b and $\bar{b}$. If the gluon lies "inside" the region of the dipole, then the angle $\phi > \pi/2$ and so $\cos\phi < 0$, which decreases the value to the likelihood here. The Higgs is most likely to radiate between the dipole by its color singlet structure, and indeed, small values of the likelihood are more Higgs-like. By contrast, if the emitted gluon lies outside the dipole region, then $\phi < \pi/2$ and so $\cos\phi > 0$, which increases the value of the likelihood. By their color octet structure, gluons are more likely to emit at wide angles, and again, large values of the likelihood are more gluon-like. Various historical jet substructure observables have been constructed to be sensitive to this specific radiation pattern as well, exhibiting much the same dependence as discussed here; see, e.g., Refs. [410, 324, 411, 412, 409].

We can estimate the range of the likelihood now with some understanding about the energy fraction of the emitted gluon, z_k. First focus on the Higgs. The Higgs is a color singlet, and so will only radiate in and around the dipole region, so that the angle between the gluon and either the b or $\bar{b}$ is of the order of the angle between the $b\bar{b}$ itself: $\theta_{bk} \sim \theta_{\bar{b}k} \sim \theta_{b\bar{b}}$. We can estimate this angle then, through the mass constraint on the Higgs:

$$m_H^2 = z_b z_{\bar{b}} E^2 \theta_{b\bar{b}}^2 \approx \frac{1}{4} E^2 \theta_{b\bar{b}}^2 \,, \tag{5.35}$$

or that

$$\theta_{b\bar{b}} \approx \frac{2m_H}{E} \,. \tag{5.36}$$

Next, for the gluon emission to still be perturbative and resolved, its relative transverse momentum to either b or $\bar{b}$ must be perturbative, above about 1 GeV, or so. Let's call this transverse momentum cut $k_\perp$. This relative transverse momentum is then

$$k_\perp \lesssim z_k E \theta_{b\bar{b}} \approx 2 z_k m_H \,, \tag{5.37}$$

or that

$$z_k \gtrsim \frac{k_\perp}{2m_H} \,. \tag{5.38}$$

Now, using this result, we can estimate the lower bound of the likelihood ratio as measured on Higgs decays:

$$\hat{\mathcal{L}}(\Phi) \gtrsim \frac{3}{4} \left(1 - \frac{\alpha_s}{2\pi} 2 C_F \frac{2m_H}{k_\perp} + \cdots \right) \,, \tag{5.39}$$

where we have used that $\cos\phi < 0$ for the region inside the dipole, and dropped constant factors for this estimate. Perturbation theory assumes an expansion about $\alpha_s = 0$ and so formally here α_s is arbitrarily close to 0. However, in practice, we set α_s to its non-zero, measured value, which in this case would be measured at the Higgs mass because this is an on-shell decay for which $\alpha_s(m_H) \approx 0.11$. Thus, the likelihood estimate can potentially become negative depending on the precise values for the cutoff scale $k_\perp$ and the numerical prefactors. Such behavior would obviously be unphysical, and is just telling us that we need even higher-order corrections to correctly describe the region of small likelihood. Nevertheless, we see that starting at next-to-leading order, there is really no non-zero lower bound on the likelihood.

71

What about the upper bound? The upper bound is set by the dynamics of the emitted gluon itself and the corresponding estimate for the energy fraction must be modified. Now, because the gluon is a color octet, the soft radiated gluon can be emitted at any angle up to the jet radius R (so we actually see the radiation in the jet). Then, from the perturbative transverse momentum cut the minimal energy fraction z_k is

$$k_\perp \lesssim z_k ER \,, \tag{5.40}$$

or that

$$z_k \gtrsim \frac{k_\perp}{ER} \,. \tag{5.41}$$

The upper bound of the likelihood ratio is therefore

$$\hat{\mathcal{L}}(\Phi) \lesssim \frac{3}{2} \left(1 + \frac{\alpha_s}{2\pi} 2C_A \frac{ER}{k_\perp} + \cdots \right) \,. \tag{5.42}$$

As energy E increases, this grows without bound, for fixed jet radius R, regardless of the value of the cutoff scale $k_\perp$. Combining these results, we therefore observe what we had anticipated from physical considerations at the beginning: at high energies, the difference in color representation of the Higgs and gluon can enable arbitrarily good discrimination, but this is only first observed at next-to-leading order.

5.3 Two-Prong Identification Without Knowledge of Mass

The fixed-order analysis of the likelihood ratio is clearly very powerful when there is both a known constraint on the mass of the signal jets as well as a fixed and known particle content, both properties of which are explicit in $H \to b\bar{b}$ identification. However, there are numerous cases, especially for new particle searches, where it is known that a signature of new physics would be massive jets, but what that mass is or what Standard Model particles the new physics decays to is unknown a priori. What can we do in that case?

Without explicit knowledge of what the new physics is, we can't just take squared matrix elements from Feynman diagrams as the probability distributions on phase space because we simply do not know them; we need another tactic. What we can do instead is to parametrize phase space as relevant for the particular signal you wish to be sensitive to by some set of IRC safe observables. Then, given that set of observables, we can determine the parametric or scaling relationships amongst them according to our expectation of signal and background jets. This technique is familiar from analyses of effective field theories and establishes a robust **power counting** that provides systematic expansion parameters about idealized limits. In this section, we will apply this power counting to the problem of discrimination of scale-invariant jets (jets with a single hard core of radiation) from jets with an intrinsic but unknown mass (jets with two hard cores of radiation). These techniques were developed and fleshed out in Ref. [413].[26]

Our first task is to determine what set of IRC safe observables we want to use to define the relevant phase space. These observables should be sufficiently expressive that they take very different values on one- or two-prong jets. We have already encountered one possible set of observables, the N-subjettinesses [360, 361, 358, 359, 362] τ_1 and τ_2 (suppressing angular exponent dependence).

---

[26]Earlier related work on power counting for multi-differential measurements are Refs. [414, 415, 416, 417].

For a jet with a single particle, $\tau_1 = \tau_2 = 0$, because there is no radiation displaced from the jet axis, while for a jet with two particles, $\tau_2 = 0$, but $\tau_1 > 0$. So these observables seem to fit the bill. However, we will instead work with the two- and three-point energy correlation functions [336, 298, 337], $e_2^{(\beta)}$ and $e_3^{(\beta)}$, defined as

$$e_2^{(\beta)} = \sum_{i<j} z_i z_j \theta_{ij}^\beta, \qquad\qquad e_3^{(\beta)} = \sum_{i<j<k} z_i z_j z_k \theta_{ij}^\beta \theta_{ik}^\beta \theta_{jk}^\beta. \qquad (5.43)$$

These observables make no reference to a jet axis and are only sensitive to products of pairwise angles between particles in the jet. As such, energy correlation functions are also recoil-free and exhibit all of the nice properties we established for quark versus gluon discrimination earlier. At any rate, analogous to τ_1 and τ_2, both $e_2^{(\beta)}$ and $e_3^{(\beta)}$ vanish on one-particle jets, while $e_3^{(\beta)} = 0$ but $e_2^{(\beta)} > 0$ on two-particle jets. So these observables will also work to parametrize the desired phase space.

As an amusing historical anecdote, the optimal observable that we will derive for this discrimination problem, called $D_2^{(\beta)}$, was (one of the many) used by experimentalists to test the most exciting result at the LHC since the discovery of the Higgs. In December 2015, both ATLAS and CMS observed an excess in the di-photon invariant mass distribution at about 750 GeV [418, 419, 420, 421]. This is the same search, though at a different mass, that was instrumental in discovering the Higgs. That both experiments observed an excess of about 3σ significance at the same mass was extremely enticing, and theorists went crazy over it, producing over 500 (!!!) papers with every possible new physics explanation presented in the following months.[27] However, as more and more data were collected, and the interesting events analyzed in new ways like testing the consistency with the hypothesis that such a new particle should produce two-prong jets, the excess shrunk and had completely disappeared by summer 2016. Such is the life of a model builder.

5.3.1 Power Counting One- Versus Two-Prong Discrimination

Let's see how this power counting approach, which focuses on the parametric scaling of emissions in the jets, proceeds. The first thing we need to do is to identify the relevant emission structure, or **modes**, in the jets of interest that we wish to discriminate. For the problem at hand, of one- versus two-prong jet discrimination, the relevant modes are illustrated in Fig. 19. Here, by "one-prong" jet, we mean, in a parametric limit, that there is a single hard core of collinear radiation c with characteristic angular size $R_{cc} \ll 1$ surrounded by wide-angle, soft radiation s of characteristic energy fraction $z_s \ll 1$. For a two-prong jet, each prong is high energy and again has characteristic angular size R_{cc}, and is surrounded by soft wide-angle radiation of energy fraction z_s. Additionally, there is a new characteristic scale corresponding to the opening angle between the two prongs, R_{12}. Radiation that fills this region about the hard prong dipole is boosted soft radiation, the collinear-soft mode cs, that has characteristic angular size R_{12} and energy fraction $z_{cs} \ll 1$. For two-prong jets to be clearly two-prongy, the angular scales must be hierarchically separated, $R_{cc} \ll R_{12} \ll 1$.

---

[27] An amusing "theory" of jumping on the bandwagon of any excesses in particle physics data was proposed in Ref. [422], that also analyzed the 750 GeV excess. For a snapshot of that time, the community response, problems with statistical interpretation, problems with citation count culture, etc., see Jester's blog post and the hundreds of comments at *Résonaances* here: `https://resonaances.blogspot.com/2016/06/game-of-thrones-750-gev-edition.html`.
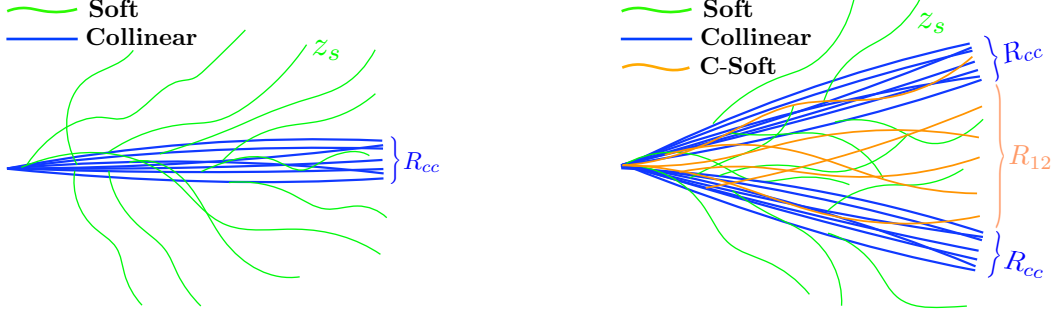
Figure 19: Illustrations of the dominant modes in one-prong (left) and two-prong (right) jets. Hard, collinear radiation is illustrated in blue, soft, wide-angle radiation in green, and for two-prong jets, soft-collinear radiation emitted from the dipole of the two prongs is orange. Figure reproduced from Ref. [413].

These modes and their corresponding energy and angular scales are presented in the table below.

Modes	Energy	Angle
Soft s	$z_s \ll 1$	1
Collinear c	1	$R_{cc} \ll R_{12} \ll 1,$
Collinear-Soft cs	$z_{cs} \ll 1$	$R_{12} \ll 1$

With these identified modes, we then use them, and only them, to evaluate the two energy correlation functions. For example, for one-prong jets, we can express its parametric contributions from soft and collinear modes by taking every possible combination of soft and collinear particles in the expression for the observables. We then plug in the energy fraction and characteristic angle into the formulas for the observables. Namely, we write

$$e_2^{(\beta)} \sim \sum_s z_s^2 + \sum_s z_s + \sum_c R_{cc}^\beta, \tag{5.44}$$

$$e_3^{(\beta)} \sim \sum_s z_s^3 + \sum_s z_s^2 + \sum_{s,c} z_s R_{cc}^\beta + \sum_c R_{cc}^{3\beta}. \tag{5.45}$$

For $e_2^{(\beta)}$, for example, the first term correlates two soft particles, the second term correlates a soft and collinear particle, and the third term correlates two collinear particles.

Next, we can drop terms that are explicitly subleading, e.g., $z_s^2 \ll z_s$. Further, because we only care about scaling relations, we will drop the explicit summation over particles in the modes, and leave that implicit. Then, on one-prong jets, the energy correlation functions have the dominant contributions from the soft and collinear modes of:

$$e_2^{(\beta)} \sim z_s + R_{cc}^\beta, \qquad\qquad e_3^{(\beta)} \sim z_s^2 + z_s R_{cc}^\beta + R_{cc}^{3\beta} \tag{5.46}$$

We can now determine the relationship between these observables on one-prong jets. If $z_s \gtrsim R_{cc}^\beta$, then $e_3^{(\beta)} \sim \left(e_2^{(\beta)}\right)^2$, because then $R_{cc}^{3\beta}$ is subleading. The other possibility is if $R_{cc}^\beta \gg z_s$ for which

we find the relationship $e_3^{(\beta)} \sim \left(e_2^{(\beta)}\right)^3$. Therefore, one-prong jets live in the region of phase space defined by the dual measurement of the energy correlation functions

$$\text{One-Prong Region: } \left(e_2^{(\beta)}\right)^3 \lesssim e_3^{(\beta)} \lesssim \left(e_2^{(\beta)}\right)^2 \tag{5.47}$$

For two-prong jets, the hard splitting itself fixes the measurement of $e_2^{(\beta)} \sim R_{12}^{\beta}$, so we only need to determine leading-power scaling for $e_3^{(\beta)}$. Also, for $e_3^{(\beta)}$ for two-prong jets, it is obvious that at leading-power we must include the correlation of the two hard cores, and then the third particle can be from anywhere else. This results in the dominant contributions to the energy correlation functions of

$$e_2^{(\beta)} \sim R_{12}^{\beta}, \qquad\qquad e_3^{(\beta)} \sim z_s R_{12}^{\beta} + z_{cs} R_{12}^{3\beta} + R_{cc}^{\beta} R_{12}^{2\beta} \tag{5.48}$$

Now, there are two possibilities to consider. First, for color-singlets, like Higgs, Z, or W bosons in the Standard Model, or, for example, electroweakinos in supersymmetry, wide-angle soft radiation exclusively comes from initial state radiation or other sources external to the physics of decay. If we assume these are negligible, by, say, going to sufficiently high boost or small jet radius, then the color-singlet two-prong region is

$$\text{Color-Singlet Two-Prong Region: } e_3^{(\beta)} \ll \left(e_2^{(\beta)}\right)^3. \tag{5.49}$$

That is, at sufficiently high boosts and small jet radii, we have $z_s \to 0$ and so wide-angle soft radiation does not contribute.

This color-singlet resonance is typically the most relevant case (and which we will focus more on below), but it is interesting to consider the case when the two-prong jet is not a color-singlet, or when there is significant contamination radiation in the jet (at high pile-up, say). In this case, the wide-angle soft radiation may be relevant, if $R_{12}^{2\beta} \lesssim z_s \ll R_{12}^{\beta}$, where the upper bound is required to define "two-prong". Because we make no more measurements and assume a relatively large jet radius, $R \sim 1$, this soft radiation actually dominates $e_3^{(\beta)}$ so that the scaling is

$$\text{Non-Color-Singlet Two-Prong Region: } e_3^{(\beta)} \ll \left(e_2^{(\beta)}\right)^2, \tag{5.50}$$

which is not parametrically different from the one-prong case. So, for identification of two-prong non-color-singlet jets, other measurements must be made to tease out individual contributions from soft, collinear, and collinear-soft radiation.

Then, for one-prong versus color-singlet two-prong jets, the parametric separatrix is the curve $e_3^{(\beta)} \sim \left(e_2^{(\beta)}\right)^3$, therefore the parametrically optimal discriminant is the ratio of this scaling, an observable called $D_2^{(\beta)}$:

$$D_2^{(\beta)} \equiv \frac{e_3^{(\beta)}}{\left(e_2^{(\beta)}\right)^3}. \tag{5.51}$$

In Fig. 20, we show plots of this phase space, with the regions dominated by one- and two-prong jets in distinct colors. At right, we also show the contours of constant values of $D_2^{(\beta)}$, which is
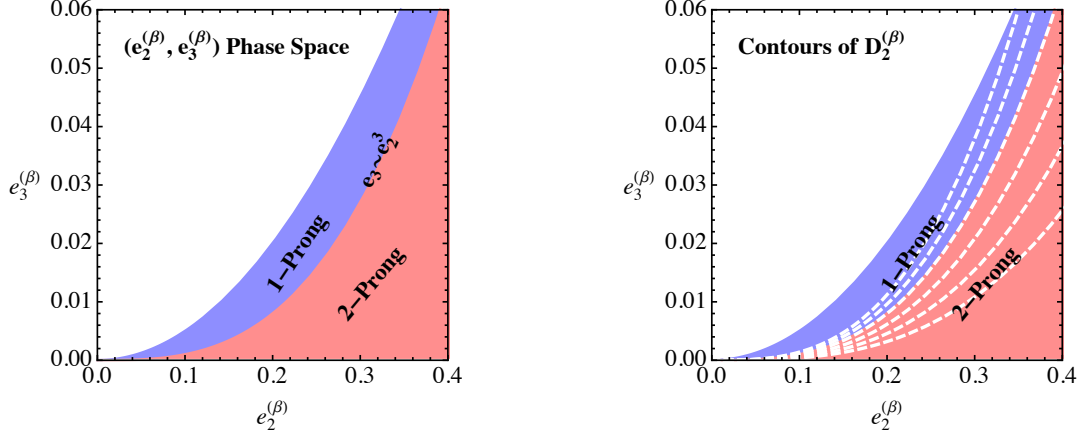
Figure 20: Plots of the phase space defined by simultaneous measurements of the energy correlation functions $e_2^{(\beta)}$ and $e_2^{(\beta)}$. Left: Illustration of the regions dominated by one-prong (purple) and two-prong jets (red). Right: Contours of the observable $D_2^{(\beta)}$ on this phase space. From Ref. [413].

the only observable whose contours do not cross the parametric scaling boundary $e_3^{(\beta)} \sim \left(e_2^{(\beta)}\right)^3$. Any other observable as a function of $e_2^{(\beta)}$ and $e_3^{(\beta)}$ will mix these regions and so a cut on another observable will necessarily have reduced discrimination power from a cut on $D_2^{(\beta)}$.

5.3.2 Boost Invariance of One- Versus Two-Prong Discrimination

There is another way to observe the optimality of $D_2^{(\beta)}$ as a discriminant for this problem. While we are working in the high-boost limit, so everything is collimated, our power counting made no specific requirements on angular scales, for example, but only needed parametric relationships. Therefore, the optimal discrimination observable for one- versus two-prong jets must necessarily be invariant to rescalings of angles in the jets that maintains the parametric relationships. However, the only physical way to rescale angles in the jets is by a Lorentz boost along the jet's direction. That is, the optimal discrimination observable separates jets into one- and two-prong classes in a Lorentz invariant way.

Let's see how this works with the functional forms of the energy correlation functions. First, we note that energy fractions z_i are invariant to boosts along the jet direction in the collimated limit because individual particle energy E_i and the total jet energy E scale in the same way:

$$z_i = \frac{E_i}{E} \rightarrow \frac{\gamma E_i}{\gamma E} = z_i \,, \tag{5.52}$$

where γ is the boost factor. By contrast, the angle between two particles i, j, θ_{ij}, scales inversely with the boost factor:[28]

$$\theta_{ij} \rightarrow \gamma^{-1} \theta_{ij} \,. \tag{5.53}$$

---

[28]This simply follows from noting that the invariant mass of two massless momenta, $2p_1 \cdot p_2 = E_1 E_2 \theta_{12}^2$ in the collinear limit, is Lorentz invariant.

With this property, we then see that under a Lorentz boost, the energy correlation functions scale as

$$e_2^{(\beta)} = \sum_{i<j} z_i z_j \theta_{ij}^\beta \to \sum_{i<j} z_i z_j \gamma^{-1} \theta_{ij}^\beta = \gamma^{-1} e_2^{(\beta)} \,, \tag{5.54}$$

$$e_3^{(\beta)} = \sum_{i<j} z_i z_j z_k \theta_{ij}^\beta \theta_{ik}^\beta \theta_{jk}^\beta \to \sum_{i<j} z_i z_j z_k \gamma^{-1} \theta_{ij}^\beta \gamma^{-1} \theta_{ik}^\beta \gamma^{-1} \theta_{jk}^\beta = \gamma^{-3} e_3^{(\beta)} \,.$$

Therefore, under a Lorentz boost along the jet direction, the ratio

$$D_2^{(\beta)} = \frac{e_3^{(\beta)}}{\left(e_2^{(\beta)}\right)^3} \to \frac{\gamma^{-3} e_3^{(\beta)}}{\left(\gamma^{-1} e_2^{(\beta)}\right)^3} = D_2^{(\beta)} \,, \tag{5.55}$$

is invariant. Another way to say this, in relationship to the phase space illustrated at right in Fig. 20, is that jets move along contours of constant $D_2^{(\beta)}$ under Lorentz boosts, never crossing the parametric separatrix.

5.3.3 Perturbative Calculation Beyond Perturbation Theory

With $D_2^{(\beta)}$ established as the likelihood ratio for optimal discrimination of one- versus two-prong jets with phase space defined by simultaneous measurement of $e_2^{(\beta)}$ and $e_3^{(\beta)}$, we would now typically move on to calculating its probability distribution on signal and background jets. The natural way to do this, within our master formula, is through the double-differential master formula

$$\frac{d^2\sigma}{de_2^{(\beta)} de_3^{(\beta)}} = \int d\Phi \, |\mathcal{M}|^2 \, \delta\left(e_2^{(\beta)} - \hat{e}_2^{(\beta)}(\Phi)\right) \delta\left(e_3^{(\beta)} - \hat{e}_3^{(\beta)}(\Phi)\right) \,. \tag{5.56}$$

Both $e_2^{(\beta)}$ and $e_3^{(\beta)}$ are IRC safe, so we can calculate this distribution order-by-order in perturbation theory to our heart's content. Then, our next step would be to marginalize this double differential distribution, slicing out contours of constant $D_2^{(\beta)}$, where

$$\frac{d\sigma}{dD_2^{(\beta)}} = \int de_2^{(\beta)} \, de_3^{(\beta)} \, \frac{d^2\sigma}{de_2^{(\beta)} de_3^{(\beta)}} \, \delta\left(D_2^{(\beta)} - \frac{e_3^{(\beta)}}{(e_2^{(\beta)})^3}\right) \,. \tag{5.57}$$

So far, this seems kosher, just following the rules of probability.

However, try as you might, this integral Eq. (5.57) in fixed-order perturbation theory does not exist [423]. We can see why simply from an analysis of the contours of $D_2^{(\beta)}$ from Fig. 20 and the definition of IRC safety. First, in the calculation of the fixed-order distribution of $D_2^{(\beta)}$, there will be a virtual contribution which lives exclusively at $D_2^{(\beta)} = 0$. This is negative and divergent, but if $D_2^{(\beta)}$ were IRC safe, then the real contribution would also be divergent at $D_2^{(\beta)} = 0$ with exactly the opposite strength, and then, when summed, divergences would cancel and all would be right with the world. However, the real contribution to $D_2^{(\beta)}$ actually diverges at *all* possible values, because all contours of $D_2^{(\beta)}$ pass through the divergent region of phase space, where $e_2^{(\beta)}, e_3^{(\beta)} \to 0$. There is simply no consistent prescription to eliminate divergences over all of phase space that should be

concentrated exclusively at $D_2^{(\beta)} = 0$. Thus, $D_2^{(\beta)}$ is not IRC safe, and so its distribution cannot be calculated order-by-order in perturbation theory.

The problem with this observable is that it is formed from the ratio of two IRC safe observables, and there is a 0/0 ambiguity if the entire phase space is allowed to contribute. A simple solution, then, is to make a cut on the denominator observable, here $e_2^{(\beta)}$, that removes the ambiguous region of phase space from consideration. However, by the approximate scale invariance of QCD at high energies, we know that the singular regions of phase space are always associated with an exponential Sudakov suppression, when all-orders effects are included. This would then seem sufficient for producing a finite distribution for $D_2^{(\beta)}$, with the caveat that we marginalized over the resummed double differential cross section:

$$\frac{d\sigma}{dD_2^{(\beta)}} = \int de_2^{(\beta)} \, de_3^{(\beta)} \, \frac{d^2\sigma^{(\text{resum})}}{de_2^{(\beta)} \, de_3^{(\beta)}} \, \delta\left(D_2^{(\beta)} - \frac{e_3^{(\beta)}}{(e_2^{(\beta)})^3}\right) . \tag{5.58}$$

Thus, a prediction for the inclusive distribution of $D_2^{(\beta)}$ cannot exist independently of resummation, and the Sudakov factor is vital for finiteness. This class of observables that are not IRC safe, but whose distribution can be rendered finite by an exponential Sudakov factor are therefore referred to as **Sudakov safe** [364, 424, 425].

We could perform this resummed calculation to study this behavior in $D_2^{(\beta)}$, but the phase space restrictions on the Lund plane imposed by measurement of $e_2^{(\beta)}$ and $e_3^{(\beta)}$ is rather complicated and the non-linear phase space constraints make bookkeeping a bit annoying. Because we already calculated it, let's instead consider calculating the N-subjettiness ratio, $\tau_{2,1}^{(\alpha)} = \tau_2^{(\alpha)}/\tau_1^{(\alpha)}$, which exhibits the same phenomena. The complete calculation for $D_2^{(\beta)}$ is presented in Ref. [425], with related calculations in Refs. [426, 103]. For the ratio observable $\tau_{2,1}^{(\alpha)}$, we had already calculated the joint resummed probability distribution of $\tau_1^{(\alpha)}$ and $\tau_2^{(\alpha)}$ in Sec. 4.3.1:

$$p_q(\tau_1^{(\alpha)}, \tau_2^{(\alpha)}) = \left(\frac{2\alpha_s C_F}{\pi\alpha}\right)^2 \frac{\log \tau_1^{(\alpha)}}{\tau_1^{(\alpha)}} \frac{\log \tau_2^{(\alpha)}}{\tau_2^{(\alpha)}} \, \exp\left[-\frac{\alpha_s}{\pi} \frac{C_F}{\alpha} \log^2 \tau_2^{(\alpha)}\right] . \tag{5.59}$$

The distribution of their ratio is then found from the marginalization integral:

$$p_q(\tau_{2,1}^{(\alpha)}) = \int d\tau_1^{(\alpha)} \, d\tau_2^{(\alpha)} \, p_q(\tau_1^{(\alpha)}, \tau_2^{(\alpha)}) \, \delta\left(\tau_{2,1}^{(\alpha)} - \frac{\tau_2^{(\alpha)}}{\tau_1^{(\alpha)}}\right) \tag{5.60}$$

$$= \sqrt{\frac{\alpha_s C_F}{\alpha}} \frac{1}{\tau_{2,1}^{(\alpha)}} \left(1 + \text{erf}\left(\sqrt{\frac{\alpha_s C_F}{\pi\alpha}} \log \tau_{2,1}^{(\alpha)}\right)\right)$$

$$= \sqrt{\frac{\alpha_s C_F}{\alpha}} \frac{1}{\tau_{2,1}^{(\alpha)}} + \frac{2\alpha_s C_F}{\pi\alpha} \frac{\log \tau_{2,1}^{(\alpha)}}{\tau_{2,1}^{(\alpha)}} - \frac{2}{3}\left(\frac{\alpha_s C_F}{\pi\alpha}\right)^2 \frac{\log^3 \tau_{2,1}^{(\alpha)}}{\tau_{2,1}^{(\alpha)}} + \mathcal{O}(\alpha_s^3) .$$

On the second line, we have written the result in closed form, in terms of an error function $\text{erf}(x)$, and then on the third line, Taylor expanded the result in powers of α_s. This makes the IRC unsafety manifest. First, in fixed-order perturbation theory, $\tau_{2,1}^{(\alpha)}$ would need two emissions in the jet to be non-zero, and so its non-trivial distribution would start at $\mathcal{O}(\alpha_s^2)$. However, the Taylor expansion of this, inclusive, distribution has contributions at orders lower than this, and further, has

78

contributions at fractional powers of α_s! By the structure and rules of Feynman diagrams, there's no way to predict especially the $\sqrt{\alpha_s}$ term, and therefore this distribution cannot be computed in fixed-order perturbation theory, and so $\tau_{2,1}^{(\alpha)}$ (and also $D_2^{(\beta)}$) is IRC unsafe, as argued.

Why the distribution starts at $\sqrt{\alpha_s}$ is rather simple to see from the structure of the Sudakov factor. The Sudakov factor has the form of a exponential distribution in $\alpha_s \log^2 \tau_2^{(\alpha)}$. Therefore, when integrated over, $\log \tau_2^{(\alpha)}$ has most of its support displaced from $\tau_2^{(\alpha)} = 0$ by an amount proportional to $1/\sqrt{\alpha_s}$, in the region where the value of the exponent is order-1.

By contrast, let's now consider bounding $\tau_1^{(\alpha)}$ from below by any non-zero $\tau_1^{(\alpha)} > \epsilon > 0$. We then find

$$p_q(\tau_{2,1}^{(\alpha)}|\tau_1^{(\alpha)} > \epsilon) = \int d\tau_1^{(\alpha)} d\tau_2^{(\alpha)} p_q(\tau_1^{(\alpha)}, \tau_2^{(\alpha)}) \delta \left(\tau_{2,1}^{(\alpha)} - \frac{\tau_2^{(\alpha)}}{\tau_1^{(\alpha)}} \right) \Theta(\tau_1^{(\alpha)} - \epsilon) \qquad (5.61)$$

$$= -\frac{2}{3} \left(\frac{\alpha_s C_F}{\pi \alpha} \right)^2 \frac{\log^2 \epsilon \left(2 \log \epsilon + 3 \log \tau_{2,1}^{(\alpha)} \right)}{\tau_{2,1}^{(\alpha)}} + \mathcal{O}(\alpha_s^3) \,.$$

As promised, this distribution starts at the right order in perturbation theory and has a Taylor expansion in α_s, and so indeed represents the distribution of an IRC safe observable. The cost of this, though, is logarithmic sensitivity to the value of the cutoff, ϵ, order-by-order, which can be large if ϵ is small.

Many examples of Sudakov safe observables are now known, with two broad classes of non-IRC safety established. The first historically (and that we discussed) are those observables that are functions of IRC safe observables and whose contours on a continuously connected set of non-zero measure of the observable pass through the singular or degenerate regions of phase space. Such observables generically have an expansion in $\sqrt{\alpha_s}$, for the reasons we discussed. A second class, and in some sense much more interesting than the first, are observables whose expansion in α_s has a non-trivial contribution at α_s^0, e.g., Refs. [427, 261, 424, 428, 429, 430, 405, 431]. The leading-order piece of such observables therefore corresponds to an ultraviolet fixed point of renormalization group evolution of QCD. By asymptotic freedom, α_s vanishes at high energies, and therefore only the leading contribution in such observables survives, resulting in a probe of universality classes of theories as defined by their non-coupling parameters (e.g., rank of gauge group, gauge representation of matter, number of quark flavors, etc.).

5.4   Linearizing the Function Basis

Let's put our machine learning hats back on and think about implementing the approaches to the problem of general resonant particle jets versus QCD jets through optimization of parameters of a neural network through gradient descent of an objective function. In this section, we were able to produce compact, functional forms for the likelihoods (and with a little more work, for the ROC curve) as functions of the phase space variables we worked with. What made this especially simple was that the dimensionality of the phase spaces were very small (at most, maybe 4 or 5) and so our brain could rather easily parse the physics content of the expressions. By contrast, if you want a machine to learn the likelihood from (simulated) data, you want that data to be representative of real physics and as such jets contain many particles with a phase space that is high dimensionality (like, say, 100). Not only can our feeble brains not imagine 100 dimensional spaces well, if you asked

the machine for the functional form of the likelihood it constructed, it would return some rather nasty function that was some linear combination of compositions of compositions of the non-linear activation functions. Just by staring at the explicit functional form of the machine's likelihood ratio, you would never know if there was some simpler approximation represented by only a few variables without significantly more work.

However, as we have emphasized here, as long as we ensure that the input data spans the data space of interest, we are free to represent the data in any way that we want, and the universal approximation theorem guarantees that the machine relaxes to the same functional form of the likelihood ratio. With a sufficiently large and general basis of functions, the likelihood can be expressed as a linear combination of additive, IRC safe observables that generalize energy correlation functions, by the Stone-Weierstrass theorem [432, 433, 434]. On a fixed manifold, this is intimately familiar from harmonic analysis in electromagnetism or quantum mechanics where, using the superposition principle, we find the complete, orthonormal basis of functions on the space of interest (spherical harmonics, Legendre polynomials, etc.). In jet physics, particle multiplicity fluctuates event-by-event and so on an ensemble of data, jets populate phase spaces of different dimensionality. To resolve this issue, one can define an overcomplete polynomial basis of IRC safe functions, referred to as energy flow polynomials [205], to accomplish this linearization. Then, one can simply use linear regression to find the coefficients of the expansion for the given task to approximate the likelihood ratio to arbitrary accuracy. What could be simpler! What's more, our brains can make sense of a power spectrum of coefficients of such a linear combination and correspondingly ascribe meaning to it, and can meaningfully filter this representation to compress the data and pull out its most important contributions. (Our brains do this all the time with Fourier series and transforms whenever we listen to music, for example.) In the soft and/or collinear limit, many of the polynomials become redundant or linearly dependent, and power counting methods can be used to establish a (less over-)complete basis [435, 436].

While we won't discuss this linear basis for likelihood estimation more, in many ways this can be the most direct way to actually figure out what the machine is learning [437, 438, 439] because the learned function through linear regression is unambiguous through (pseudo)inversion of a linear equation, and the corresponding calculation is straightforward (if tedious when phase space is high-dimensional) because all observables are themselves IRC safe.

Exercises

1. We discussed in detail the fixed-order analysis of the IRC safe discrimination problem $H \to b\bar{b}$ versus $g \to b\bar{b}$. After $H \to b\bar{b}$ and $H \to WW^*$ (W^* = an off-shell W boson), the third most common decay mode of the Higgs is $H \to gg$, which happens about 10% of the time. However, unlike the dominant hadronic decay mode $H \to b\bar{b}$, there is little hope to observe the $H \to gg$ channel. With our understanding of the IRC safe likelihood ratio, can we help improve the situation at all? Consider the binary discrimination problem $H \to gg$ versus $g \to gg$ in highly-boosted, collimated jets. What are the leading and next-to-leading order likelihood ratios now? How is this different than $H \to b\bar{b}$ identification? What challenges with $H \to gg$ identification do you envision and are they surmountable?

2. Let's consider one- versus two-prong discrimination again, but let's now define phase space by simultaneous measurement of $\tau_1^{(\alpha)}$ and $\tau_2^{(\alpha)}$. Perform the corresponding power counting analysis like we did with the energy correlation functions. What is the expression for the

likelihood ratio on this phase space now? Where do signal and background events live? Have we seen this observable before?

3. As stated in the beginning of this section, one of the goals of jet substructure (at least historically, e.g., Refs. [366, 410, 440, 441, 442, 358, 443]) was the identification of boosted, hadronically-decaying top quarks. At leading order, top quarks decay to three massless partons through the weak interaction, and in general, all three partons carry an order-1 energy fraction and are all a similar angle from one another. Additionally, the top quark has a fixed mass, $m_t \approx 173$ GeV. Therefore, background jets initiated by light QCD partons must also have this mass and as such must exhibit a strong two-prong structure. Therefore, top versus QCD discrimination is in general a two- versus three-prong jet discrimination problem, on jets with a fixed mass.

   (a) Perform the power counting analysis for this problem on the phase space defined by simultaneous measurement of $\tau_2^{(\alpha)}$ and 3-subjettiness, $\tau_3^{(\alpha)}$. What is the resulting likelihood observable you find? Where on this phase space do signal and background events live?

   (b) Perform the power counting analysis for this problem on the phase space defined by the simultaneous measurement of the two- and three-point correlation functions, $e_2^{(\beta)}$ and $e_3^{(\beta)}$, and the four-point correlation function $e_4^{(\beta)}$, where

   $$e_4^{(\beta)} = \sum_{i<j<k<l} z_i z_j z_k z_l \theta_{ij}^{\beta} \theta_{ik}^{\beta} \theta_{il}^{\beta} \theta_{jk}^{\beta} \theta_{jl}^{\beta} \theta_{kl}^{\beta} \,. \tag{5.62}$$

   What is the likelihood in terms of these observables? Where on phase space do signal and background events live? See Ref. [444] for more details.

6 Third Example: Quark vs. Gluon Jet Discrimination Redux

In the past two lectures, we had established a couple of properties of the likelihood ratio for different discrimination problems in QCD that were not evident a priori. First, in our analysis of quark versus gluon discrimination, even from the simple Lund plane picture of scale invariant soft and collinear emissions, we found that sensitivity to multiple emissions improved discrimination power. Second, in our analysis of one- versus two-prong discrimination useful in new physics searches, we found that the likelihood ratio was not IRC safe, even though it was constructed as a function of IRC safe observables. Putting these things together, perhaps a fixed-order or even a perturbative analysis is insufficient for actually establishing the form of the likelihood ratio for general binary discrimination problems in QCD. In this lecture, we will revisit quark versus gluon discrimination, with a reconsideration of the Lund plane and work to identify the likelihood ratio for emissions on any subset of the plane with finite area. This work will suggest that total hadronic multiplicity is actually *the* likelihood ratio for quark versus gluon discrimination, which is not an IRC safe observable. So, we will have to stretch and strengthen our tools for going outside the friendly confines of Feynman diagrams and our master formula. This will prepare us for the final lecture in which we cast off any pretense of a perturbative analysis whatsoever.

6.1 Likelihood on the Lund Plane and Perturbative Multiplicity

Let's go back to the Lund plane picture of scale-invariant emissions in a jet and see if, simply from this and the difference in color factors C_F versus C_A, we can identify the likelihood observable for quark versus gluon discrimination. To ensure that intermediate steps of this argument are always finite, let's consider some finite area region of the Lund plane as our region of interest; we only consider emissions in that region as contributing to the likelihood, and all emissions outside of that region are ignored. An example of this is illustrated in Fig. 21. Recall that in this simple picture, the density of emissions on the Lund plane is uniform, with

$$\rho_q \propto \frac{2\alpha_s}{\pi} C_F \,, \qquad\qquad \rho_g \propto \frac{2\alpha_s}{\pi} C_A \,, \qquad\qquad (6.1)$$

the emission densities of quark and gluon jets, respectively. Because the emission density is uniform, the likelihood observable is independent of the area of interest on the Lund plane and correspondingly must be independent of the particular energy fractions z_i and splitting angles θ_i of the emissions. Therefore, the only remaining piece of information about the jets that can possibly be used for discrimination on the Lund plane is the number or **multiplicity** of emissions in the region of interest. The emission multiplicity is therefore the optimal quark versus gluon discriminant within the Lund plane approximation [445, 446]. For now, we will only consider emissions in the so-called "primary" Lund plane, corresponding to direct emission from the hard, initiating particle, ignoring secondary (or later) emissions from already emitted particles.

Let's make this argument that the multiplicity of emissions on the Lund plane is the likelihood ratio more precise. Given emissions on the Lund plane, the joint probability distribution of anything we could possibly measure on it is

$$p(\{z_i, \theta_i\}_{i=1}^n, n) = p(\{z_i, \theta_i\}_{i=1}^n | n)\, p(n) = p(n) \prod_{i=1}^n p(z_i, \theta_i) \,, \qquad (6.2)$$

where n is the multiplicity of emissions and $\{z_i, \theta_i\}$ are the energy fraction and angle of emission i with respect to the jet axis. In the middle, we have simply used the definition of conditional probability and at right, we note that all emissions in the Lund plane are independent, so the energy fraction and angles of distinct emissions are all uncorrelated. Given a finite area of interest on the Lund plane $\mathcal{A}$, the probability distributions of the energy fractions and angles are

$$p(z, \theta) = \frac{\int_{\mathcal{A}} \frac{dz'}{z'} \frac{d\theta'}{\theta'}\, \delta(z - z')\delta(\theta - \theta')}{\int_{\mathcal{A}} \frac{dz'}{z'} \frac{d\theta'}{\theta'}} \,. \qquad\qquad (6.3)$$

Because this is a normalized distribution, there is no sensitivity to the color Casimir of the jet and this distribution is not only identical for all emissions, but insensitive to the initiating particle of the jet, as well. Now, with this construction, we simply take the ratio of this multi-differential distribution on gluon and quark jets to produce the likelihood:

$$\mathcal{L} = \frac{p_g(\{z_i, \theta_i\}_{i=1}^n, n)}{p_q(\{z_i, \theta_i\}_{i=1}^n, n)} = \frac{p_g(n) \prod_{i=1}^n p(z_i, \theta_i)}{p_q(n) \prod_{i=1}^n p(z_i, \theta_i)} = \frac{p_g(n)}{p_q(n)} \,, \qquad (6.4)$$

exclusively dependent on the multiplicity n, with no dependence on the kinematics of the emissions in the jet whatsoever.
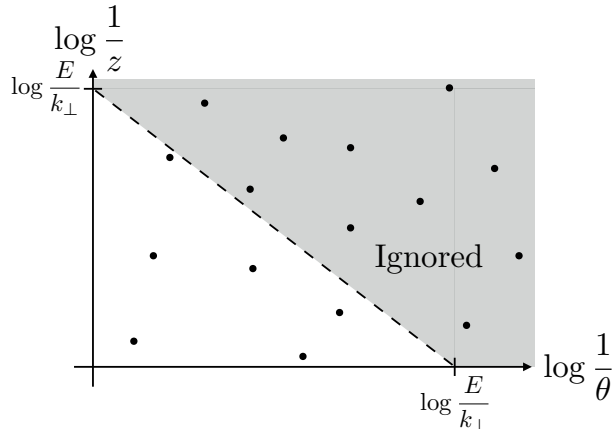
Figure 21: Region of interest on the (primary) Lund plane of emissions. Here, emissions are ignored if their relative transverse momentum $k_\perp$ to the jet axis is below a cut value.

Now, we see the necessity of the restriction to a finite area on the Lund plane. Assuming true scale invariance, of course any jet will emit an infinite number of particles, extending to arbitrarily low scales. QCD, of course, is not really a conformal or scale-invariant field theory, but is only approximately at relatively large energies, energies well above the scale at which the coupling becomes large and where hadron masses are, at around 1 GeV. So, for our perturbative analysis to make sense and to indeed be a decent approximation, we should restrict emissions on the Lund plane to lie in the perturbative region. Emissions become non-perturbative, sensitive to hadronic as opposed to partonic physics, when the relative transverse momentum of the particle to the jet axis falls below about 1 GeV. In general, however, we will introduce a transverse momentum cut $k_\perp$ and demand that the region of interest on the Lund plane are those emissions with relative transverse momentum above this cut. This region of interest is shown in Fig. 21 for jets with energy E. Note that the cut is a diagonal line through the relationship of the Lund plane coordinates and the transverse momentum of

$$k_\perp = zE\theta. \tag{6.5}$$

6.1.1 Derivation of Poissonian Distribution

Calculating the distribution of emission multiplicity is simple using the facts we already know about the Lund plane. Emissions in this region of interest are independent of one another, with constant probability proportional to the density from Eq. (6.1). We could just skip to the answer, but, as was the case with the Sudakov factor, the derivation of the distribution of multiplicity is illuminating, so we will proceed slowly. Just like we did for the Sudakov factor in Sec. 4.2.2, we will split up this region into N equal area pieces, and then take the $N \to \infty$ limit. Let's assume that we find k emissions in the region of Fig. 21 on a quark jet. That is, k of the N total pieces have emissions, while $N - k$ have no emissions, and the area of each small piece is

$$A_N = \frac{\log^2 \frac{E}{k_\perp}}{2N}. \tag{6.6}$$

83

That is, the emission probability into any one piece is the area times the density:

$$p_{\text{emit}} = A_N \rho_q = \frac{\alpha_s C_F}{\pi} \frac{\log^2 \frac{E}{k_\perp}}{N} \, . \tag{6.7}$$

Then, the probability for there to be k emissions in the region of interest $p_q(k)$ is

$$p_q(k) \propto p_{\text{emit}}^k (1 - p_{\text{emit}})^{N-k} = \left(\frac{\alpha_s C_F}{\pi} \frac{\log^2 \frac{E}{k_\perp}}{N} \right)^k \left(1 - \frac{\alpha_s C_F}{\pi} \frac{\log^2 \frac{E}{k_\perp}}{N} \right)^{N-k} \, . \tag{6.8}$$

With all N pieces of the region of interest identical, these k emissions can of course be located anywhere, so we need a binomial factor to sum over all possible emission combinatorics:

$$p_q(k) = \binom{N}{k} \left(\frac{\alpha_s C_F}{\pi} \frac{\log^2 \frac{E}{k_\perp}}{N} \right)^k \left(1 - \frac{\alpha_s C_F}{\pi} \frac{\log^2 \frac{E}{k_\perp}}{N} \right)^{N-k} \, . \tag{6.9}$$

Now, we can take the $N \to \infty$ limit of this expression, with fixed value for k, so that $k/N \to 0$. In this limit, we see that the no emission factor exponentiates, and the binomial factor reduces to

$$\lim_{N \to \infty} \binom{N}{k} \frac{1}{N^k} = \frac{N^k}{k!} \frac{1}{N^k} = \frac{1}{k!} \, . \tag{6.10}$$

Then, putting these pieces together, the probability that there are k emissions in a quark jet's perturbative region of interest is Poissonian distributed with

$$p_q(k) = \frac{1}{k!} \left(\frac{\alpha_s C_F}{\pi} \log^2 \frac{E}{k_\perp} \right)^k \exp \left[-\frac{\alpha_s C_F}{\pi} \log^2 \frac{E}{k_\perp} \right] \, . \tag{6.11}$$

For both quark and gluon jets, this can be expressed compactly as

$$p_q(k) = \frac{\lambda_q^k e^{-\lambda_q}}{k!} \, , \qquad\qquad p_g(k) = \frac{\lambda_g^k e^{-\lambda_g}}{k!} \, , \tag{6.12}$$

where λ_q and λ_g are the mean multiplicities in the perturbative region where,

$$\lambda_q = \frac{\alpha_s C_F}{\pi} \log^2 \frac{k_\perp}{E} \, , \qquad\qquad \lambda_g = \frac{\alpha_s C_A}{\pi} \log^2 \frac{k_\perp}{E} \, . \tag{6.13}$$

These Poissonian distributions are very interesting and of a completely different form than that of the Sudakov distributions we had derived for angularities, for example. In particular, as Poissonian distributions, their means are equal to their variances, and so both exhibit Casimir scaling:

$$\frac{\langle n_q \rangle}{\langle n_g \rangle} = \frac{C_F}{C_A} \, , \qquad\qquad \frac{\sigma_q^2}{\sigma_g^2} = \frac{C_F}{C_A} \, , \tag{6.14}$$

where $\langle n_q \rangle$ is the mean quark jet multiplicity on the Lund plane and σ_q^2 is its variance. By contrast, no such simple Casimir scaling relationship exists for moments of Sudakov distributions (which you can verify!), but rather the cumulative distributions of the angularities are related by
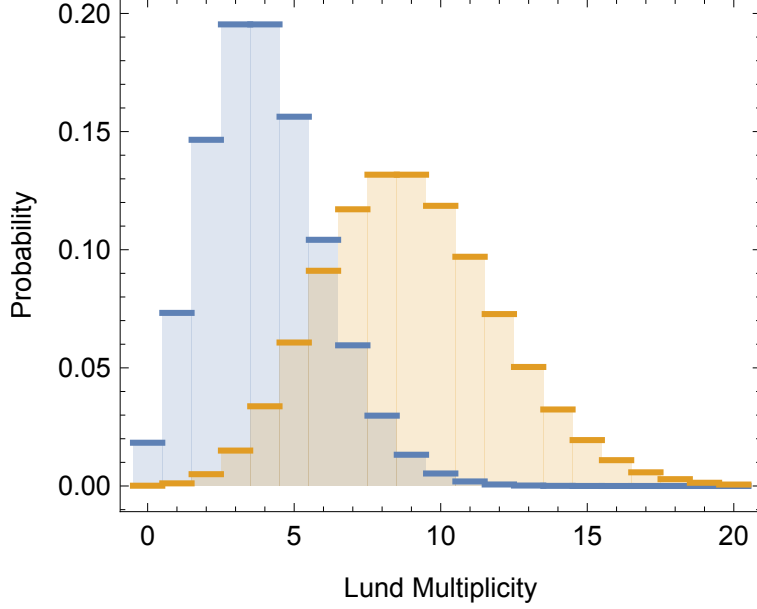
Figure 22: Plots of the Poisson distributions of the Lund plane multiplicity for quark jets (blue) and gluon jets (orange). For this plot, the mean values $\lambda_q = 4$ and $\lambda_g = 9$ were used, which have the correct ratio as expected from Casimir scaling.

Casimir scaling. Further, this comparison between the Sudakov distribution of angularities and these Lund plane multiplicity distributions is a bit unfair, because we made no such perturbative cut on emissions for the angularities. If one does this, to have an apples-to-apples comparison of discrimination power, the story for angularities changes significantly, which you will study in the exercises.

A useful measure of discrimination power that we will employ in other cases when cumulative distributions are a bit unwieldy is the ratio of the difference of means to the square root of the sum of variances:

$$\eta \equiv \frac{|\mu_q - \mu_g|}{\sqrt{\sigma_q^2 + \sigma_g^2}} \, . \tag{6.15}$$

We have illustrated example Poisson distributions in Fig. 22, where the means (and variances) satisfy Casimir scaling in QCD, Eq. (6.14). The sum of the variances sets a ruler or natural distance scale between the distributions, and so this measure is the distance between distributions in these natural units. If this distance is small, then the distributions have large overlap and hence weak discrimination power, and conversely, if the distance is large then the overlap is small, and discrimination power is strong. For Poissonian Lund plane multiplicity, we have

$$\eta = \frac{|\langle n_q \rangle - \langle n_g \rangle|}{\sqrt{\sigma_q^2 + \sigma_g^2}} = \frac{\lambda_g - \lambda_q}{\sqrt{\lambda_q + \lambda_g}} = \sqrt{\frac{\alpha_s}{\pi}} \frac{C_A - C_F}{\sqrt{C_F + C_A}} \log \frac{E}{k_\perp} \, . \tag{6.16}$$

The scale of the cutoff $k_\perp$ is fixed above the mass scale of the hadrons, $k_\perp \sim 1$ GeV, while the energy E of the jet is something we can control. As we increase the energy of the jet, this distance

85

between the distributions grows without bound, and so discrimination power can become arbitrarily good at high energies.

This observation is rather similar to testing the fairness of a flipped coin as the number of flips grows. As the energy or mean multiplicity increases, the distributions become more and more narrow as compared to their mean values. Further, gluon jets will have approximately twice as many emissions as quark jets on the Lund plane. Therefore, if you observe twice as many emissions as you expect for a quark jet, the probability that the jet was actually a quark jet vanishes extremely quickly with mean multiplicity.

As another interesting measure of discrimination power, let's look at the reducibility factors, and the maximal purity of quark and gluon jets with this Lund plane multiplicity. The likelihood ratio of the Poisson distributions is

$$\mathcal{L} = \frac{p_g(k)}{p_q(k)} = \left(\frac{\lambda_g}{\lambda_q}\right)^k e^{-(\lambda_g - \lambda_q)} = \left(\frac{C_A}{C_F}\right)^k e^{-\frac{\alpha_s}{\pi}(C_A - C_F)\log^2 \frac{E}{k_\perp}} . \tag{6.17}$$

The maximal purity of a quark jet sample with this observable is the minimal value of the likelihood, where $k = 0$:

$$\mathcal{L}_{\min} = e^{-\frac{\alpha_s}{\pi}(C_A - C_F)\log^2 \frac{E}{k_\perp}} . \tag{6.18}$$

Interestingly, and unlike the case for angularities, a pure sample of quark jets is not possible with the Lund plane multiplicity. Nevertheless, this minimal value is exponentially suppressed in the large jet energy limit, so the gluon contamination to a pure quark sample is practically very small. By contrast, the maximum gluon jet purity is determined by the maximal value of the likelihood ratio, but because $C_A > C_F$, this is divergent:

$$\mathcal{L}_{\max} = \lim_{k \to \infty} \left(\frac{C_A}{C_F}\right)^k e^{-\frac{\alpha_s}{\pi}(C_A - C_F)\log^2 \frac{E}{k_\perp}} \to \infty . \tag{6.19}$$

Thus, a pure sample of gluons is possible with the Lund plane multiplicity, again, unlike the situation with measurement of an angularity.

6.1.2 Large Multiplicity Approximation

From these expressions for the Poisson distributions of the Lund plane multiplicity, we could calculate ROC curves, AUCs, etc., but the discrete nature of the Poisson distribution renders these expressions complicated and unilluminating. So, here we will continue and make an additional approximation in which the average number of emissions is very large, $\lambda_q, \lambda_g \gg 1$, for which the distributions can be very well approximated by a continuous distribution. Again, we could just write down the answer where the Poissonians transform into Gaussians, but the derivation is illuminating in itself, so we provide the details here.

Just focusing on the multiplicity distribution for quark jets, we can express this distribution as

$$p_q(k) = \frac{\lambda_q^k e^{-\lambda_q}}{k!} = e^{-\lambda_q + k \log \lambda_q - \log k!} , \tag{6.20}$$

putting everything in the exponent. This is an exact relationship, but this exponential form will make approximation in the $\lambda_q \to \infty$ limit straightforward. As an exponential, this distribution will

be dominated by the region about where its argument takes its maximal value. Expanding about this maximum is called the **saddlepoint approximation**, which here is effectively a version of the central limit theorem. Note that if $k \ll \lambda_q$, then the first term in the exponent dominates, and the distribution is exponentially suppressed. Therefore, the multiplicity k itself must be large, $k \gg 1$, and so we can expand the factorial with Stirling's approximation. The leading contributions from Stirling's approximation for the asymptotic behavior of the factorial are

$$\lim_{k \to \infty} \log k! = k \log k - k + \mathcal{O}(\log k) \,. \tag{6.21}$$

Using this approximation, the exponent can be expressed as

$$-\lambda_q + k \log \lambda_q - \log k! \approx -\lambda_q + k \log \lambda_q - k \log k + k \,. \tag{6.22}$$

Its maximum, or stationary point, is where its derivative vanishes:

$$\frac{d}{dk} \left(-\lambda_q + k \log \lambda_q - k \log k + k \right) = \log \lambda_q - \log k = 0 \,, \tag{6.23}$$

or that the extremum is where $k_0 = \lambda_q$. We can then approximate the exponent as a quadratic function near its maximum by taking a second derivative, where

$$\frac{d^2}{dk^2} \left(-\lambda_q + k \log \lambda_q - k \log k + k \right) \bigg|_{k=\lambda_q} = -\frac{1}{\lambda_q} \,. \tag{6.24}$$

Then, the Taylor expansion through quadratic order of the exponent is

$$-\lambda_q + k \log \lambda_q - \log k! \approx -\frac{(k - \lambda_q)^2}{2\lambda_q} + \cdots \,. \tag{6.25}$$

Then the Poisson probability distribution in this large mean and continuous limit becomes

$$p_q(x) \propto e^{-\frac{(x - \lambda_q)^2}{2\lambda_q}} \,, \tag{6.26}$$

which is a Gaussian. We only write proportional here, because we need to fix up the normalization, but that is easy:

$$p_q(x) = \frac{1}{\sqrt{2\pi\lambda_q}} e^{-\frac{(x - \lambda_q)^2}{2\lambda_q}} \,, \qquad\qquad p_g(x) = \frac{1}{\sqrt{2\pi\lambda_g}} e^{-\frac{(x - \lambda_g)^2}{2\lambda_g}} \,, \tag{6.27}$$

writing both the quark and gluon large multiplicity (or, correspondingly, high energy) limiting distributions.

The likelihood ratio in this continuous emission limit takes the form

$$\mathcal{L} = \frac{p_g(x)}{p_q(x)} = \sqrt{\frac{\lambda_q}{\lambda_g}} \exp \left[-\frac{\lambda_g - \lambda_q}{2\lambda_g\lambda_q} (x^2 - \lambda_g\lambda_q) \right] \,, \tag{6.28}$$

which is (still!) monotonic in the multiplicity x, and so x itself is the optimal discrimination observable. We can correspondingly directly calculate the AUC as our familiar ordered integral, where

$$\text{AUC} = \int dx \, dy \, p_q(x) \, p_g(y) \, \Theta(x - y) \tag{6.29}$$

$$= \int dx\,dy\, \frac{1}{\sqrt{2\pi\lambda_q}}\, e^{-\frac{(x-\lambda_q)^2}{2\lambda_q}}\, \frac{1}{\sqrt{2\pi\lambda_g}}\, e^{-\frac{(y-\lambda_g)^2}{2\lambda_g}}\, \Theta(x-y)$$

$$= \frac{1}{2}\left(1 - \mathrm{erf}\left(\frac{\lambda_g - \lambda_q}{\sqrt{2}\sqrt{\lambda_q + \lambda_g}}\right)\right)\,.$$

Fascinatingly, the argument of the error function is exactly the discrimination distance measure we defined in Eq. (6.15) (which justifies it a posteriori). As the means of the distributions get large, $\lambda_q, \lambda_g \to \infty$, the argument of the error function also diverges, and the AUC vanishes, because $\mathrm{erf}(\infty) = 1$. Thus, again we see that perfect discrimination power is possible in this limit.

6.1.3 Lund Plane Multiplicity with Running Coupling

This analysis of the Lund plane multiplicity can be extended to higher orders, including the effects of physics that start only formally beyond the simple double logarithmic limit in which we have worked. Many more details and calculations are presented in the literature, e.g., Refs. [446, 312, 447, 448], but here we will just discuss one effect, the running, or scale-dependence, of the coupling α_s. QCD exhibits asymptotic freedom in which its coupling α_s decreases in value as the relevant energy scale increases. As a consequence, because emission rates are proportional to the value of the coupling, fewer particles are emitted at higher energies as compared to lower energies. This has the effect of increasing the density of emissions on the Lund plane as you move away from the origin, toward low scales off at infinity. Let's calculate the effect of this variation of emission density.

First, we need to establish the scale dependence of the coupling itself. This is governed by the β-function of QCD, where the dependence of α_s on the energy scale μ is

$$\mu\frac{\partial \alpha_s}{\partial \mu} = \beta(\alpha_s) = -\frac{\beta_0}{2\pi}\alpha_s^2 + \cdots\,, \tag{6.30}$$

only keeping terms to leading-logarithmic or one-loop accuracy in the β-function. The coefficient β_0 is [449, 450] (see also Ref. [451, 452])

$$\beta_0 = \frac{11}{3}C_A - \frac{4}{3}n_f T_R\,, \tag{6.31}$$

which depends on the number of active quark flavors, n_f. Actually, we have already seen this before, in Sec. 4.3.2, as the contribution to the distribution of angularities from hard, collinear emissions in a gluon jet. The solution to the β-function equation with one-loop running is

$$\alpha_s(\mu) = \frac{2\pi}{\beta_0\,\log\frac{\mu}{\Lambda_{\mathrm{QCD}}}}\,, \tag{6.32}$$

where Λ_{QCD} is the scale at which the value of the (perturbative) coupling diverges. This scale can be determined by matching the value of α_s at a fixed scale, typically taken to be the mass of the Z boson, and accounting for the variation of the number of active quarks as the scale μ increases (especially the effect of the bottom quark must be accounted for). Experimental numbers can be taken from the PDG [453], but we won't worry about that here, and just use this functional form for our analysis.

Even though the value of the coupling varies with scale, and therefore location on the Lund plane, at this approximation, emissions are still independent and so the resulting multiplicity distribution is still Poissonian. Therefore, all we need to calculate is the mean value of the multiplicity

with this running coupling. This calculation is now a non-trivial integral, and not simply proportional to the area of the region of interest, but is nevertheless not complicated to evaluate. For quark jets, for example, the mean multiplicity with running coupling is then

$$\lambda_q = \frac{2C_F}{\pi} \int \frac{dz}{z} \frac{d\theta}{\theta} \, \alpha_s(z\theta E) \, \Theta\left(z\theta - \frac{k_\perp}{E}\right) \tag{6.33}$$

$$= \frac{4C_F}{\beta_0} \log \frac{E}{\Lambda_{\text{QCD}}} \left(\log \frac{\log \frac{E}{\Lambda_{\text{QCD}}}}{\log \frac{k_\perp}{\Lambda_{\text{QCD}}}} - 1\right) + \frac{4C_F}{\beta_0} \log \frac{k_\perp}{\Lambda_{\text{QCD}}} \,,$$

where the scale at which the coupling is evaluated is the relative transverse momentum of the emission, $z\theta E$. The mean Lund plane multiplicity on gluon jets is simply related by replacing $C_F \to C_A$, as usual, so the distributions still exhibit Casimir scaling of their moments. As energy increases, this mean multiplicity increases more slowly than the double logarithmic result of Eq. (6.13). The leading contribution in the $E \to \infty$ limit, with fixed Λ_{QCD} and $k_\perp$, scales like

$$\lim_{E \to \infty} \lambda_q \to \frac{4C_F}{\beta_0} \log \frac{E}{\Lambda_{\text{QCD}}} \log \log \frac{E}{\Lambda_{\text{QCD}}} \,, \tag{6.34}$$

which has an interesting logarithm of a logarithm structure. Nevertheless, this still diverges as $E \to \infty$, and so perfect discrimination is still possible in this limit.

6.2 Baby Steps Beyond Perturbation Theory: Discrimination with Hadronic Multiplicity

This analysis that simply the number of emissions in the primary Lund plane is the likelihood observable, at least to this approximation, suggests that the particle multiplicity in jets in general is either the likelihood observable, or is, at least, a very good discriminant (if not technically optimal). However, once we open up to an observable that simply counts the number of particles in the jet, we lose IRC safety. The argument that multiplicity is not IRC safe is very simple; consider an initial jet with n particles. Now, emit an exactly 0 energy gluon, so that the total number of particles in the jet is now $n + 1$, but there is no way to observe a 0 energy gluon. Thus, the value of the multiplicity changes from emission of 0 energy gluons, and so it cannot be IRC safe. The solution we had arrived at earlier was to cut off the Lund plane or require a minimal resolution for any emission that contributed to the multiplicity. However, we now want to remove that restriction, counting *all* particles in the jet, which suffers from IRC unsafety.

It needs to be emphasized that a lack of IRC safety does not place multiplicity (or any IRC unsafe observable) into second-class, or that somehow IRC safe observables are "better" than their non-IRC safe counterparts. We had already seen in the case of Sudakov safe observables, which are not IRC safe, a way to calculate them outside of fixed-order perturbation theory. Similarly, for multiplicity, we will have to broaden our scope and possible tools for understanding it, and can't rely on our Fermi's Golden Rule master formula with Feynman diagrams as the be-all, end-all of a theoretical analysis. This illustrates the flexibility necessary for working with (more) realistic jets: we have to let Nature and the rules of statistics determine optimal observables for discrimination, and not be biased by our theoretical prejudices.

For the remainder of this lecture, we will study the total multiplicity of hadrons inside jets and introduce techniques and approximations that have been identified as useful starting point
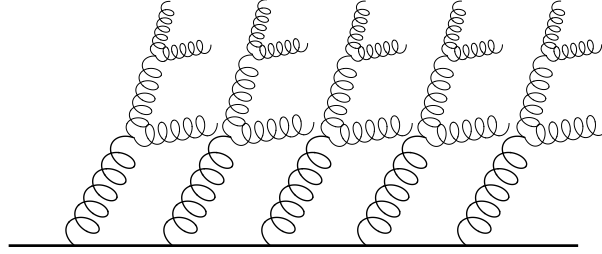
Figure 23: Schematic picture of dominant perturbative emission structure. Gluons are emitted off of the hard, initiating particle (solid line) with a rate proportional to its quadratic color Casimir. Then, those secondary gluons emit further gluons that emit further gluons in a quasi-scale invariant process.

approximations when Feynman rules are of limited use. We will see that scale dependence of the mean hadronic multiplicity can be calculated perturbatively, but for discrimination, we need the full functional form of the distribution, not just its mean. The key observation that allows us to make progress is that of KNO scaling, where the functional form of the multiplicity distribution seems to be universal, and just rescaled by the mean multiplicity. This then connects hadronic multiplicity distributions at various energy scales with no additional parameters. To end this section, we compare the discrimination power ROC curves for quark versus gluon discrimination for the various observables we have considered throughout these lectures.[29]

6.2.1 Scale Dependence of Mean Multiplicity

To calculate the mean hadronic multiplicity, we need to establish a number of assumptions, and then we can get to calculating. First things first: let's establish a dominant picture for the total partonic multiplicity. This is illustrated in Fig. 23. The picture we have in our mind is as follows. First, gluons are emitted off of the initiating, hard particle in proportion to its quadratic color Casimir, just like in our picture of the Lund plane. This is the dominant contribution for primary emissions because these gluons exhibit both soft and collinear divergences. Now, with gluons emitted, the dominant source of secondary, tertiary, or higher emissions will be more gluons from these primary gluons for the two reasons that gluons have a larger color factor than quarks and again, the squared matrix element for emitted gluons exhibits both soft and collinear singularities. The number of emissions off of the initiating particle is effectively fixed, and so the total multiplicity of partons is dominantly gluons and dominantly from gluons that emit gluons that emit gluons that emit gluons....

We don't want the partonic multiplicity, we want the hadronic multiplicity, so we need to connect this picture with that of measured hadrons. We will make the assumption, which has some theoretical justification in a phenomenon called **preconfinement** [454, 455, 456, 457], that after this showering process, once the partons have a relative scale of about 1 GeV from one another,

---

[29]Consideration of and bias to IRC safe observables is absolutely justified for precision studies, in which a systematically-improvable perturbative calculation is compared to experimental data. In such a case, it is necessary to quantify theoretical uncertainties. For our study of optimal discrimination observables and likelihoods outside of perturbation theory, our goal is markedly distinct: we want to establish robust correlations and scaling relationships between observables that can inform a more principled machine learning study.

they arrange themselves into color singlet clusters that independently fragment or "hadronize" into a fixed number of hadrons. This clustering and fragmentation process therefore multiplies the partonic multiplicity by a constant, fixed factor which we cannot predict, but can nevertheless extract from experimental data. That is, if we can derive a homogeneous differential equation for the scale dependence of mean multiplicity, this overall constant factor will not appear, and we will only need a description of the generation of perturbative, partonic multiplicity.

To derive such a differential equation, we need two remaining assumptions. Because we want to determine the scale μ dependence of the mean hadronic multiplicity $\langle n \rangle$, the differential equation will be in the variable μ, which correspondingly orders the emission structure. So, we will need to determine a reasonable (and defensible!) ordering prescription to the generation of emissions. To determine this, we look to results in electromagnetism first. In the study of high-energy cosmic ray showers, it was observed that ionization radiation from electron-position pairs was emitted at subsequently smaller angles as the shower progressed, a phenomenon now referred to as the **Chudakov effect** [458, 459]. This is a consequence of quantum coherence of emission radiation in electromagnetism, and suppresses radiation emitted at wider angles than that of the initial dipole. An exactly analogous phenomenon exists in QCD through the quantum coherence of color, and soft gluon emission is suppressed at angles wider than that of their emitting dipole. The Chudakov effect or color coherence then implies that emissions are ordered in the scale set by their relative angle, $\mu = \theta E$, for a jet with energy E.

This property of color coherence or angular ordering can be observed by azimuthal averaging of the eikonal emission factor of a gluon k off of a dipole. Recall that the eikonal emission factor from a $q\bar{q}$ dipole is, for example,

$$S_{q\bar{q}}(k) \propto \frac{s_{q\bar{q}}}{s_{qk}s_{k\bar{q}}} \ . \tag{6.35}$$

This expression is independent of the energy of the q or $\bar{q}$ in the dipole, and the energy of the gluon k is just an overall factor. For azimuthal averaging, we can ignore that energy dependence. In the collinear limit, then, the angular dependence in the eikonal factor is

$$S_{q\bar{q}}(k) \propto \frac{\theta_{q\bar{q}}^2}{\theta_{qk}^2 \theta_{k\bar{q}}^2} = \frac{\theta_{q\bar{q}}^2 + \theta_{k\bar{q}}^2 - \theta_{qk}^2}{2\theta_{qk}^2 \theta_{k\bar{q}}^2} + \frac{\theta_{q\bar{q}}^2 + \theta_{qk}^2 - \theta_{k\bar{q}}^2}{2\theta_{qk}^2 \theta_{k\bar{q}}^2} \ . \tag{6.36}$$

At right, we have separated the eikonal factor into two terms, each of which has a single collinear divergence. For example, the expression

$$\frac{\theta_{q\bar{q}}^2 + \theta_{k\bar{q}}^2 - \theta_{qk}^2}{2\theta_{qk}^2 \theta_{k\bar{q}}^2} \tag{6.37}$$

has a collinear divergence when $k \parallel q$, but no divergence if $k \parallel \bar{q}$. Thus, this representation of the eikonal factor separates it into two distinct sectors, where in each we can identify a unique emitter, at least in the well-defined collinear limit.

What makes this sectoring interesting is that angular ordering is manifest in each sector separately. To see this, let's use the law of cosines to re-express the non-collinear angle $\theta_{k\bar{q}}^2$ as an azimuthal angle, and write

$$\frac{\theta_{q\bar{q}}^2 + \theta_{k\bar{q}}^2 - \theta_{qk}^2}{2\theta_{qk}^2 \theta_{k\bar{q}}^2} = \frac{\theta_{q\bar{q}}^2 - \theta_{q\bar{q}}\theta_{qk}\cos\phi}{\theta_{qk}^2 \left(\theta_{q\bar{q}}^2 + \theta_{qk}^2 - 2\theta_{q\bar{q}}\theta_{qk}\cos\phi \right)} \tag{6.38}$$

$$= \frac{\theta_{q\bar{q}}^2(\theta_{q\bar{q}}^2 + \theta_{qk}^2) - 2\theta_{q\bar{q}}^2\theta_{qk}^2 \cos^2 \phi}{\theta_{qk}^2 \left((\theta_{q\bar{q}}^2 + \theta_{qk}^2)^2 - 4\theta_{q\bar{q}}^2\theta_{qk}^2 \cos^2 \phi\right)} + \text{odd in } \phi.$$

In the second line, we have expressed the denominator as a manifestly even function on $\phi \in [0, \pi]$, and in the numerator have ignored terms odd in ϕ on that same domain. Then, azimuthal averaging this is easy:

$$\int_0^{2\pi} \frac{d\phi}{2\pi} \frac{\theta_{q\bar{q}}^2 + \theta_{k\bar{q}}^2 - \theta_{qk}^2}{2\theta_{qk}^2\theta_{k\bar{q}}^2} = \int_0^\pi \frac{d\phi}{\pi} \frac{\theta_{q\bar{q}}^2(\theta_{q\bar{q}}^2 + \theta_{qk}^2) - 2\theta_{q\bar{q}}^2\theta_{qk}^2 \cos^2 \phi}{\theta_{qk}^2 \left((\theta_{q\bar{q}}^2 + \theta_{qk}^2)^2 - 4\theta_{q\bar{q}}^2\theta_{qk}^2 \cos^2 \phi\right)} \tag{6.39}$$

$$= \frac{1}{\theta_{qk}^2} \Theta(\theta_{q\bar{q}}^2 - \theta_{qk}^2).$$

That is, after azimuthal averaging, the sectored eikonal factor consists exclusively of the collinearly divergent angular factor, and further, is only non-zero if the emission angle is smaller than the opening angle of the dipole.

The final piece of the puzzle is the relevant scale for particle production. For a particle to be produced, that is, to actually *observe* that a particle has been emitted, that particle must have a sufficiently large relative transverse momentum to its emitter. Note that the relative transverse momentum of a particle with energy fraction z and splitting angle θ is $k_\perp = z\theta E = z\mu$, where μ is the angular ordering scale established earlier. Finally, these considerations motivate the following homogeneous evolution equation for the mean multiplicity:

$$\langle n(\mu + \delta\mu) \rangle = \langle n(\mu) \rangle + \frac{\delta\mu}{\mu} \frac{2\alpha_s C_A}{\pi} \int_0^1 \frac{dz}{z} \langle n(z\mu) \rangle. \tag{6.40}$$

Here, $\delta\mu$ is a small change in the evolution scale μ. As μ increases, two things can happen: either nothing (first term on the right), or a soft gluon can be emitted at a relative transverse momentum of $z\mu$, and we need to sum over all possible emission energies, consistent with the scale μ. By taking $\delta\mu \to 0$, this can equivalently be expressed as an integro-differential equation:

$$\mu \frac{\partial}{\partial\mu} \langle n(\mu) \rangle = \frac{2\alpha_s C_A}{\pi} \int_0^1 \frac{dz}{z} \langle n(z\mu) \rangle. \tag{6.41}$$

To solve this, we make a power-law ansatz, where

$$\langle n(\mu) \rangle \propto \mu^\gamma, \tag{6.42}$$

where γ is called the **anomalous dimension**. This equation then becomes

$$\gamma = \frac{2\alpha_s C_A}{\pi} \int_0^1 dz\, z^{-1+\gamma} = \frac{2\alpha_s C_A}{\pi} \frac{1}{\gamma}. \tag{6.43}$$

The anomalous dimension is therefore

$$\gamma = \sqrt{\frac{2\alpha_s C_A}{\pi}}. \tag{6.44}$$

This dependence on $\sqrt{\alpha_s}$ can be explained as the fractal dimension [460, 461] of the shape of the multiparticle Lund planes produced as the shower progresses [462]. The solution to this evolution equation with fixed coupling in a truly scale-invariant theory is then

$$\langle n(\mu) \rangle \propto \mu^{\sqrt{\frac{2\alpha_s C_A}{\pi}}} . \tag{6.45}$$

That is, the mean hadronic multiplicity in a jet, in the extreme high energy limit that QCD is scale-invariant, scales with a fractional power of the energy scale μ of the jet. Note that this grows faster with μ than any power of $\log \mu$ and so the total multiplicity increases much faster with energy than the primary Lund multiplicity.

This description of the mean multiplicity can be improved through incorporation of higher-order effects. With this result and one-loop running of the coupling, note that the mean multiplicity satisfies the differential equation

$$\mu \frac{\partial}{\partial \mu} \langle n(\mu) \rangle = \sqrt{\frac{2\alpha_s C_A}{\pi}} \langle n(\mu) \rangle = \sqrt{\frac{4 C_A}{\beta_0 \log \frac{\mu}{\Lambda_{\text{QCD}}}}} \langle n(\mu) \rangle . \tag{6.46}$$

The solution is

$$\langle n(\mu) \rangle \propto e^{\sqrt{\frac{16 C_A}{\beta_0} \log \frac{\mu}{\Lambda_{\text{QCD}}}}} = e^{\sqrt{\frac{32\pi C_A}{\beta_0^2} \frac{1}{\alpha_s(\mu)}}} , \tag{6.47}$$

where at right we have replaced the explicit scale dependence with the value of the running coupling, $\alpha_s(\mu)$. More details about this calculation and how higher-order effects can be included can be found in the original literature [463, 464, 465, 466, 467, 468, 447].

From our stated assumptions, we expect that the mean multiplicity is proportional to the appropriate quadratic color Casimir of the initiating particle; $\langle n_q \rangle \propto C_F$ for quark jets and $\langle n_g \rangle \propto C_A$ for gluon jets. However, this intuition comes from a perturbative analysis and understanding, and so the only limit in which a perturbative description dominates is in the extreme high-energy limit, $\mu \to \infty$. The approach to true Casimir scaling of the mean multiplicities as a function of the jet scale μ is obviously very important for discrimination power of multiplicity, and is experimentally observed to be extremely slow. Recall that in QCD, $C_A/C_F = 2.25$, but the observed ratio of average multiplicity of gluon to quark jets at energies of 200 GeV is about 1.5, while even at an energy of 1.2 TeV is still only about 1.7 [469]. This data from ATLAS is plotted in Fig. 24 and compared with both simulated data and the calculated energy dependence of the mean multiplicity, as in Eq. (6.47), though to much higher perturbative accuracy. We will see in Sec. 6.3 that there is strong sensitivity of discrimination power to the assumed relationship between quark and gluon jet multiplicity.

6.2.2 Higher Moments of Multiplicity and KNO Scaling

One can continue this analysis of the moments of the hadronic multiplicity distribution, and derive differential equations for their scale dependence. These results can then be re-interpreted as the scale dependence of parameters in the complete multiplicity distribution itself, and so one can observe and predict how the distribution is warped and molded by physics at different energy scales. Here, instead of this, we will listen to the experimentalists' results for the measurement of these higher moments in collision events at various energies. Even by the late 1960s, it was observed
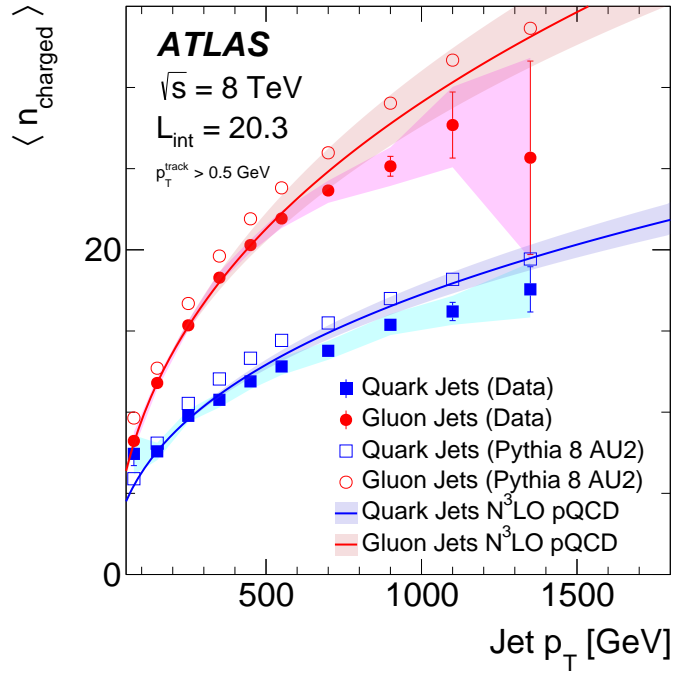
Figure 24: Plot of the mean charged particle multiplicity on quark (blue) and gluon (red) jets as measured at the ATLAS experiment, as a function of jet energy. Data (solid dots) are compared to simulated data (empty dots), and to the theoretical calculation (solid curves), an extension of Eq. (6.47) including three higher orders in the perturbative expansion of the anomalous dimension. From Ref. [469].

in data that the second moment of the hadronic multiplicity was proportional to the square of the mean:

$$\langle n^2 \rangle \propto \langle n \rangle^2 \,. \tag{6.48}$$

Even higher moments were measured, and the pattern continued, with the kth moment of the multiplicity distribution proportional to the kth power of the mean:

$$\langle n^k \rangle \propto \langle n \rangle^k \,. \tag{6.49}$$

In the early 1970s, this was formalized through a simple scaling relationship of the hadronic multiplicity distribution, $p(n)$, where

$$p(n) = \frac{1}{\langle n \rangle} \, \psi \left(\frac{n}{\langle n \rangle} \right) \,, \tag{6.50}$$

where $\psi(x)$ is some universal, normalized, distribution with unit mean. This relationship is called **KNO scaling** [470, 471].[30]

---

[30]It was actually first established by Polyakov, but was named after Koba, Nielsen, and Olesen. This is merely one of an infinity of examples of Stigler's law of eponymy [472] in particle physics.

It is important to note that KNO scaling is a property of the total hadronic multiplicity, and is not apparent in different definitions of (sub)multiplicities. For instance, the Poisson distribution of the primary Lund plane multiplicity that we studied earlier this lecture clearly does not satisfy KNO scaling. The mean and variance of the Poisson distribution are equal, $\mu = \sigma^2$, but KNO scaling would predict that the variance is proportional to the mean squared, $\sigma^2 \propto \mu^2$. The original justification for KNO scaling followed from Bjorken and Feynman scaling [473, 474, 475], early identified properties of the strong interactions (before the development of QCD) that are a consequence of its scale invariance.[31]

The minimal assumption from which KNO scaling is a consequence is that the multiplicity distribution is invariant to scaling of jet energies. By "scale-invariant" we mean that the multiplicity distribution at a given energy scale E, $p(n|E)$, transforms simply when energy is scaled by $\lambda > 0$:

$$p(n|E) = \lambda^{-\gamma}\, p(\lambda^\gamma n|\lambda E)\,. \tag{6.51}$$

That is, as the energy is scaled by λ, the multiplicity n is correspondingly scaled by λ^γ.[32] Thus, the relationship between the multiplicity and energy must take the form

$$p(n|E) = \frac{1}{kE^\gamma}\, \psi\left(\frac{n}{kE^\gamma}\right)\,, \tag{6.52}$$

where $\psi(x)$ is some energy-independent function and k is a constant. Note that the mean multiplicity from this scale-invariant assumption is

$$\langle n \rangle = \int dn\, n\, p(n|E) = \int dn\, \frac{n}{kE^\gamma}\, \psi\left(\frac{n}{kE^\gamma}\right) = kE^\gamma \int dx\, x\, \psi\left(x\right) = kE^\gamma\,. \tag{6.53}$$

In the penultimate expression, we can set the mean value of $\psi(x)$ to 1 with appropriate normalization because it is energy-independent. As demonstrated in Eq. (6.45), the mean multiplicity does indeed satisfy this fractional power dependence on energy, in the limit that QCD is scale-invariant. Therefore, we indeed have that

$$p(n) = \frac{1}{\langle n \rangle}\, \psi\left(\frac{n}{\langle n \rangle}\right)\,, \tag{6.54}$$

which is KNO scaling. The particular form of the KNO function $\psi(x)$, as well as violations of KNO scaling, can be predicted in perturbative QCD, see, e.g., Refs. [476, 463, 477, 478, 479, 480, 481, 482, 483, 484, 485, 468]. While we will present a specific parametrization of the KNO function shortly, in the analysis of these notes, however, we will mostly only need the existence of KNO scaling.

KNO scaling can be directly observed by comparing multiplicity data from experiments at different center-of-mass collision energies, for example. At left in Fig. 25 are the total charged hadronic multiplicity distributions in e^+e^- collision events at four different center-of-mass collision energies [486, 487, 488]. Naturally, as the collision energy increases, so too does the multiplicity, as follows from the result in Eq. (6.47). However, if each of these distributions is rescaled by

---

[31]My Ph.D. advisor Michael Peskin made the defensible claim that James "BJ" Bjorken was the first person to understand conformal field theories.

[32]Of course, this relationship isn't perfect because the multiplicity n is integral valued and so arbitrary scale transformations are not allowed, but we can assume that the support of $p(n|E)$ is over a sufficiently large domain that the continuous and smooth approximation is valid.
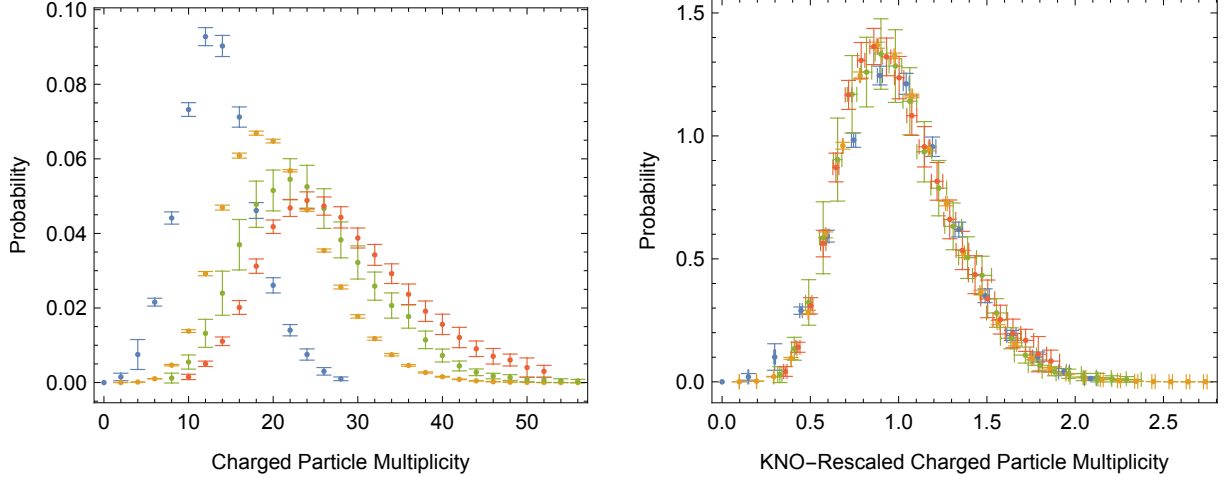
Figure 25: Charged particle multiplicity distributions (left) in $e^+e^- \to$ hadrons collisions at $\sqrt{s} = 34, 91, 161, 206$ GeV, and the same distributions rescaled by their means (right). Data from Refs. [486, 487, 488].

its respective mean, as prescribed by KNO scaling, the rescaled multiplicity distributions snap into very nearly the exact same functional distribution, as illustrated at right in Fig. 25. One small thing to note is that these experimental data only correspond to observed charged particles because individual charged particles can be observed by their ionization track in the detector. To good approximation, however, the relationship between charged and all hadronic particles is just an overall constant factor, which follows from approximate isospin conservation. We will say much more about this in the next lecture.

6.2.3 Discrimination from KNO Scaling of Multiplicity

With the assumption of KNO scaling of the hadronic multiplicity distribution, let's now work to calculate its discrimination power. In this section, we will work to establish a bound on the discrimination power as defined through the AUC, and in the next section, use a convenient parametrization from which we can calculate the full ROC curve. For quark versus gluon discrimination with hadronic multiplicity, the AUC can be calculated from

$$\text{AUC} = \int dn_q \, dn_g \, p_q(n_q) \, p_g(n_g) \, \Theta(n_q - n_g) \,, \tag{6.55}$$

where $p_q(n)$ and $p_g(n)$ are the quark and gluon multiplicity distributions, respectively. Let's now assume KNO scaling for the multiplicity distributions, with

$$p_q(n) = \frac{1}{\langle n_q \rangle} \psi\left(\frac{n}{\langle n_q \rangle}\right) \,, \qquad\qquad p_g(n) = \frac{1}{\langle n_g \rangle} \psi\left(\frac{n}{\langle n_q \rangle}\right) \,. \tag{6.56}$$

Not only do we assume KNO scaling of the distributions, but that the universal distribution $\psi(x)$ is identical for quarks and gluons. This seems like a reasonable assumption, and consistent with all of the results we have derived thus far, but is potentially interesting to test and understand

96

its implications in more detail. In Sec. 6.3.2, we will test this hypothesis from a couple different experimental analyses.

With these KNO scaled distributions, the AUC can be written as

$$\text{AUC} = \int dn_q \, dn_g \, \frac{1}{\langle n_q \rangle} \psi \left(\frac{n_q}{\langle n_q \rangle} \right) \frac{1}{\langle n_g \rangle} \psi \left(\frac{n_g}{\langle n_g \rangle} \right) \Theta(n_q - n_g) \tag{6.57}$$

$$= \int dx \, dy \, \psi(x) \, \psi(y) \, \Theta \left(\frac{\langle n_q \rangle}{\langle n_g \rangle} x - y \right)$$

$$= \int dx \, \psi(x) \, \Psi \left(\frac{\langle n_q \rangle}{\langle n_g \rangle} x \right) .$$

Here, we have introduced the cumulative KNO function

$$\Psi(x) = \int_0^x dx' \, \psi(x') . \tag{6.58}$$

With Casimir scaling, the ratio of mean multiplicities is set by the ratio of quadratic Casimirs,

$$\frac{\langle n_q \rangle}{\langle n_g \rangle} = \frac{C_F}{C_A} , \tag{6.59}$$

but even relaxing Casimir scaling, we still expect that $\langle n_q \rangle < \langle n_g \rangle$, and so the AUC is necessarily less than 1/2, indicating useful discrimination power. If $\langle n_q \rangle = \langle n_g \rangle$, the AUC would be 1/2 (the multiplicity distributions would be identical), so let's see if we can provide an estimate for the AUC in terms of an expansion about the point $\langle n_q \rangle = \langle n_g \rangle$.

Let's introduce the deviation from unit ratio $\mathcal{R}_{qg}$, where

$$\frac{\langle n_q \rangle}{\langle n_g \rangle} = 1 - \mathcal{R}_{qg} , \tag{6.60}$$

and $0 < \mathcal{R}_{qg} < 1$ because $0 < \langle n_q \rangle < \langle n_g \rangle$. Just writing out the first few orders in the expansion explicitly, the AUC can be expressed as

$$\text{AUC} = \int dx \, \psi(x) \, \Psi \left((1 - \mathcal{R}_{qg}) \, x \right) \tag{6.61}$$

$$= \int dx \, \psi(x) \, \Psi(x) - \mathcal{R}_{qg} \int dx \, x \, \psi(x)^2 + \frac{\mathcal{R}_{qg}^2}{2} \int dx \, x^2 \, \psi(x) \, \psi'(x) + \cdots$$

$$= \frac{1}{2} - \left(\mathcal{R}_{qg} + \frac{\mathcal{R}_{qg}^2}{2} \right) \int dx \, x \, \psi(x)^2 + \cdots .$$

In going from the second to third lines, we used integration by parts, and assumed that the KNO distribution vanished at the endpoint $x = 0$. The remaining integral of the KNO function expanded to this order,

$$\int dx \, x \, \psi(x)^2 , \tag{6.62}$$

has no upper bound (which can easily be seen by taking the extreme distribution $\psi(x) = \delta(x - 1)$), but does have a non-trivial lower bound.

To establish this lower bound, let's collect properties and assumptions of the KNO function. We have and assume:

97

1. The KNO function is non-negative on $x \in [0, \infty)$ and is normalized:

$$1 = \int_0^\infty dx \, \psi(x) \,. \tag{6.63}$$

2. The mean value of the KNO function is 1:

$$1 = \int_0^\infty dx \, x \, \psi(x) \,. \tag{6.64}$$

3. We assume that the probability for a multiplicity of 0 is 0:

$$\psi(0) = 0 \,. \tag{6.65}$$

4. Because the KNO function is observable, all of its moments exist:

$$\int_0^\infty dx \, x^n \, \psi(x) < \infty \,, \qquad \forall \, n > 0 \,. \tag{6.66}$$

To establish the lower bound,[33] we can use the Cauchy-Schwarz inequality, for which we have

$$\left(\int dx \, e^{-tx} \, x \, \psi(x) \right)^2 \leq \left(\int dx \, x \, \psi(x)^2 \right) \left(\int dx \, x \, e^{-2tx} \right) \,, \tag{6.67}$$

where we have introduced the exponential as a helper function and $t > 0$ is some parameter. Rearranging and evaluating the exponential integral, we have a bound on the integral of interest:

$$4t^2 \left(\frac{d}{dt} \int dx \, e^{-tx} \, \psi(x) \right)^2 \leq \int dx \, x \, \psi(x)^2 \,, \tag{6.68}$$

In this way, the integral in the parentheses is the KNO function's Laplace transform:

$$\mathcal{L}[\psi](t) = \int dx \, e^{-tx} \, \psi(x) \,, \tag{6.69}$$

and so the inequality is

$$4t^2 \left(\frac{d}{dt} \mathcal{L}[\psi](t) \right)^2 \leq \int dx \, x \, \psi(x)^2 \,. \tag{6.70}$$

We would like to establish the greatest lower bound, and so let's maximize with respect to t. Differentiating the left side and setting it to 0, we find

$$\frac{d}{dt} \left[4t^2 \left(\frac{d}{dt} \mathcal{L}[\psi](t) \right)^2 \right] = 0 \,, \tag{6.71}$$

---

[33]This is the best argument for a maximum lower bound I could construct. If you have a better approach with fewer assumptions, I would be greatly interested!

or that

$$\left(\frac{d}{dt} + t\,\frac{d^2}{dt^2}\right)\mathcal{L}[\psi](t) = 0\,. \tag{6.72}$$

Because we assume that all moments of the KNO distribution $\psi(x)$ exist, the Taylor expansion of its Laplace transform is meaningful and converges. Through t^2, the Taylor expansion of the Laplace transform is

$$\mathcal{L}[\psi](t) = 1 - t + \frac{\langle x^2\rangle t^2}{2} + \cdots = e^{-t}\left(1 + \frac{\sigma^2 t^2}{2} + \cdots\right)\,. \tag{6.73}$$

On the right, we have just extracted the overall exponential suppression, which modifies the coefficients of the expansion from moments about 0 to central moments, about the mean. By the third and fourth assumptions, the central moment expansion of the Laplace transform rapidly converges because the KNO function is necessarily rather narrowly distributed about its mean, $\sigma^2 \lesssim 1$. So, we can safely ignore higher-order contributions in the central moment expansion. Terminating the raw moment expansion at low orders is much less justified.

Plugging this expansion through the variance term into the extremization formula, Eq. (6.72), we find that the value of t that maximizes the lower bound is $t = 1$, independent of the value of the variance, σ^2. Then, we can evaluate the lower bound at $t = 1$:

$$4t^2\left(\frac{d}{dt}\mathcal{L}[\psi](t)\right)^2\bigg|_{t=1} = 4t^2\left(\frac{d}{dt}e^{-t}\left(1 + \frac{\sigma^2 t^2}{2} + \cdots\right)\right)^2\bigg|_{t=1} \tag{6.74}$$

$$= \frac{4}{e^2}\left(1 - \frac{\sigma^2}{2} + \cdots\right)^2\,.$$

This is generically non-zero. Further, with our assumptions, the variance of the KNO distribution is bounded from above by $\sigma^2 \lesssim 1$, and so we can plug in $\sigma^2 = 1$ to relax the bound a bit more:

$$\frac{4}{e^2}\left(1 - \frac{\sigma^2}{2} + \cdots\right)^2\bigg|_{\sigma^2=1} \approx \frac{1}{e^2} \approx 0.135\,. \tag{6.75}$$

Now, returning way back to the estimate of the AUC and putting these results together, we have

$$\text{AUC} = \frac{1}{2} - \left(\mathcal{R}_{qg} + \frac{\mathcal{R}_{qg}^2}{2}\right)\int dx\,x\,\psi(x)^2 + \cdots \lesssim \frac{1}{2} - \frac{1}{e^2}\left(\mathcal{R}_{qg} + \frac{\mathcal{R}_{qg}^2}{2}\right)\,. \tag{6.76}$$

For relatively small $\mathcal{R}_{qg} = 1 - \langle n_q\rangle/\langle n_g\rangle$, $\mathcal{R}_{qg} \lesssim 0.7$ (so the term in parentheses with explicit $\mathcal{R}_{qg}$ is less than 1), this should be a rather robust upper bound on the AUC.

6.2.4 An Explicit Parametrized KNO Function

Just using abstract KNO scaling and the few other assumptions we made about the properties of the KNO function, we can only provide so much detail as to the discrimination power of the hadronic multiplicity distribution. Essentially since the time that KNO scaling was identified, groups have presented numerous parametrized expressions for the KNO distribution (e.g., Refs. [464, 463, 489,

490, 491, 492, 493, 494, 495, 496, 497, 498, 499]), and here we will study one such representation. By our assumption that all moments of the KNO distribution exist, the KNO distribution must vanish as $x \to \infty$ more rapidly than any power of x. This is easily accomplished by assuming that the KNO distribution has an exponential tail:

$$\psi(x) \to e^{-ax} , \tag{6.77}$$

for $a > 0$, as $x \to \infty$. Further, we also assume that the KNO distribution vanishes at $x = 0$, which is easily accomplished with a non-negative power of x near $x = 0$:

$$\psi(x) \to x^b , \tag{6.78}$$

for $b > 0$, as $x \to 0$. Then, along with the normalization and mean restrictions, this suggests the one-parameter form

$$\psi(x|k) = \frac{k^k}{\Gamma(k)} \, x^{k-1} e^{-kx} , \tag{6.79}$$

where $k > 1$. This distribution is normalized and has unit mean:

$$1 = \int_0^\infty dx \, \psi(x|k) = \int_0^\infty dx \, x \, \psi(x|k) \tag{6.80}$$

Its variance is

$$\sigma^2 = \int_0^\infty dx \, (x-1)^2 \, \psi(x|k) = \frac{1}{k} , \tag{6.81}$$

which is indeed less than 1, for $k > 1$. For the data in Fig. 25, the KNO function is fit extremely well to this function with $k = 10$.[34]

    This functional form of the KNO distribution is the continuous generalization of the negative binomial distribution. The negative binomial distribution can be constructed from a mixture of a Gamma distribution and a Poisson distribution in the following way. The mean of the Poisson distribution λ can be considered a random variable itself, distributed according to a Gamma distribution. The Gamma distribution is the maximal entropy probability distribution with fixed, positive mean and fixed expectation value of the logarithm of the random variable. Thus, the interpretation of the negative binomial distribution for multiplicity is that the particle production is independent and uniform over a volume of phase space that fluctuates with maximal entropy. Planck was apparently the first to identify the negative binomial as the distribution of bosons among k clusters with equal probability [496, 510, 511].

---

[34]This particular parametrized functional form for the KNO function is also often the same that is used to parametrize non-perturbative contributions to additive IRC safe observables and is called a **shape function** [500, 501, 502, 503, 504, 505]. In that context, the non-perturbative shape function is convolved with the perturbative differential cross section to produce a distribution that can describe data over all of phase space. Further, the low moments of the non-perturbative corrections for many observables are known to take a universal form [506, 507, 508, 509]. There may be a rather direct connection between KNO scaling and shape functions as perturbative calculations are accurate where resolved particle multiplicity is small (just a few hard emissions), while non-perturbative physics dominates when multiplicity is large (every pion is resolved), but I am unaware of such an analysis in the literature.

With the negative binomial assumption for the KNO distribution, we can then calculate the corresponding AUC for quark versus gluon discrimination. From the general analysis in Eq. (6.57), the AUC can be expressed as

$$\text{AUC} = \int dx\, dy\, \psi(x|k)\, \psi(y|k)\, \Theta\left(\langle n_q \rangle x - \langle n_g \rangle y\right) \tag{6.82}$$

$$= \frac{k^{2k}}{\Gamma(k)^2} \int dx\, dy\, x^{k-1} y^{k-1} e^{-k(x+y)}\, \Theta\left(\langle n_q \rangle x - \langle n_g \rangle y\right)$$

$$= 1 - \frac{\Gamma(2k)}{k\,\Gamma(k)^2} \left(\frac{\langle n_g \rangle}{\langle n_q \rangle}\right)^k \left(1 + \frac{\langle n_g \rangle}{\langle n_q \rangle}\right)^{-2k} {}_2F_1\left(2k, 1, 1+k, \frac{\langle n_g \rangle}{\langle n_g \rangle + \langle n_q \rangle}\right).$$

Here, ${}_2F_1(a, b, c, z)$ is the hypergeometric function. This particular form isn't so illuminating, but we can express $\langle n_q \rangle = (1 - \mathcal{R}_{qg})\langle n_g \rangle$ and Taylor expand in $\mathcal{R}_{qg}$ to produce

$$\text{AUC} = \frac{1}{2} - \frac{4^{-k}\Gamma(2k)}{\Gamma(k)^2}\left(\mathcal{R}_{qg} + \frac{\mathcal{R}_{qg}^2}{2}\right) + \cdots . \tag{6.83}$$

The form of this expansion of course follows the general results we had established with the bounds that we derived in the previous section. In particular, we can validate the bound, by saturating the Gamma function coefficient at $k = 1$, the minimum value allowed by our assumptions. At this point, note that

$$\left.\frac{4^{-k}\Gamma(2k)}{\Gamma(k)^2}\right|_{k=1} = \frac{1}{4} > \frac{1}{e^2}, \tag{6.84}$$

and so the AUC of the negative binomial distribution is indeed less than our bound. As k increases, the coefficient scales like $\sqrt{k}$, which can be verified through Stirling's formula,

$$\frac{4^{-k}\Gamma(2k)}{\Gamma(k)^2} = \frac{\sqrt{k}}{2\sqrt{\pi}} + \mathcal{O}(k^{-1/2}) \cdots . \tag{6.85}$$

With this explicit parametrization of the multiplicity distribution, we can also determine the reducibility factors for samples of maximal quark and gluon purity. The likelihood ratio of the quark and gluon multiplicity distributions is

$$\mathcal{L} = \frac{p_q(n)}{p_g(n)} = \frac{\langle n_g \rangle\, \psi\left(\frac{n}{\langle n_q \rangle}\big|k\right)}{\langle n_q \rangle\, \psi\left(\frac{n}{\langle n_g \rangle}\big|k\right)} = \left(\frac{\langle n_g \rangle}{\langle n_q \rangle}\right)^k e^{-kn\frac{\langle n_g \rangle - \langle n_q \rangle}{\langle n_q \rangle \langle n_g \rangle}}. \tag{6.86}$$

From this, we immediately see that a pure gluon sample can be attained with a cut at large multiplicity,

$$\lim_{n \to \infty} \mathcal{L} = 0 \tag{6.87}$$

A pure quark sample is technically not possible, as the maximal value of the likelihood is limited by the parameter k:

$$\lim_{n \to 0} \mathcal{L} = \left(\frac{\langle n_g \rangle}{\langle n_q \rangle}\right)^k = e^{-k\log\frac{\langle n_q \rangle}{\langle n_g \rangle}}. \tag{6.88}$$
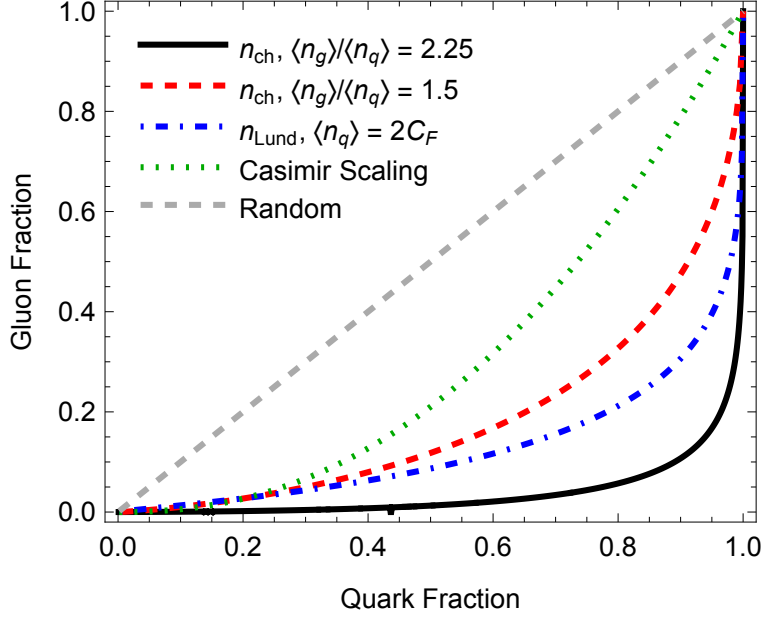
Figure 26: Comparison of ROC curves of the classes of quark versus gluon discriminators considered in these lectures: Casimir scaled additive IRC safe observables (green dotted), primary Lund plane multiplicity with $\langle n_q \rangle = 2C_F$ (blue dot-dashed), total (charged) hadronic multiplicity with $\langle n_g \rangle / \langle n_q \rangle = 1.5$ (red dashed), and total (charged) hadronic multiplicity with $\langle n_g \rangle / \langle n_q \rangle = C_A / C_F = 2.25$ (black). For the total charged multiplicity curves, we use the negative binomial distribution with $k = 10$. Dashed gray is the random classifier as reference.

In practice, however, we demonstrated that k is large, $k \sim 10$, and even if the ratio of means is much weaker than Casimir scaling, say only $\langle n_g \rangle / \langle n_q \rangle \sim 1.5$, the maximum value of the likelihood still exceeds 50. We note that this pattern of reducibility factors is similar to what we observed with Lund plane multiplicity as well, that a pure gluon sample was attainable, but a quark sample will always be contaminated by an exponentially small amount of gluon jets.

6.3 Comparison of Three Techniques for Quark vs. Gluon Discrimination

With the collected results for quark versus gluon discrimination presented throughout these lectures, it is useful to compare them all, as quantified by their ROC curves. The comparison of Casimir scaling in additive IRC safe observables (like angularities), Lund plane multiplicity, and total charged particle multiplicity is displayed in Fig. 26. The ROC curve for IRC safe angularities has no parameters (it is fixed by the ratio C_A / C_F), while the multiplicity distributions have parameters that fix the mean multiplicities. For the primary Lund plane multiplicity, we take $\langle n_q \rangle = 2C_F$, which, from Eq. (6.33) for the mean multiplicity with one-loop running of the coupling, corresponds to jets with an energy of about $E \sim 200$ GeV. For the total multiplicity curves, we fix the form of the KNO distribution to be the negative binomial with $k = 10$, but choose two different values for the ratio of quark and gluon expectation values, $\langle n_g \rangle / \langle n_q \rangle = 1.5$ and $= C_A / C_F = 2.25$.

These plots make it clear that observables like the angularities that are dominantly sensitive to

a single emission in the jet are not optimal for this discrimination problem. Multiplicity observables are much more powerful discriminants, and their discrimination power improves as the ratio of the means of gluon to quark jets increases. Sensitivity to total multiplicity with full infinite energy Casimir scaling of means (solid black in Fig. 26) produces extremely good discrimination, with an AUC of $\text{AUC} \approx 0.04$ (from Eq. (6.82)).

6.3.1 Mutual Irreducibility and Quarks and Gluons from Mixed Samples

As discussed all the way back in the beginning of Sec. 4, practically defining the ground truth labels of "quark" and "gluon" jets has migrated to the rather lazy and theoretically imprecise practice of whatever jet is asked for in selecting the process of interest in an event simulator. Through our analysis, and especially through the identification of the multiplicity as such a powerful discriminant, another technique suggests itself. Namely, quark and gluon jet labels are not set a priori, but rather are an output from modeling mixed samples as some linear combination of two "topics", which we can, a posteriori, call "quark" and "gluon", if so desired. This technique of jet **topic modeling** [512, 105, 106] enables extraction of individual quark and gluon jet distributions from unlabeled data, as long as those data satisfy a few reasonable requirements.

First, the number of linearly-independent samples must be equal to the number of topics that one desires to model. For quark and gluon jets, this means that we need two samples of jets and further, the relative quark and gluon fractions in those samples must be different (but whose precise values do not need to be known). For example, one might choose the leading jet from, say, $pp \to Z+$ jet and $pp \to$ dijets samples. Second, the likelihood ratio observable measured on those samples of jets must be mutually irreducible, which means that pure samples of the two topics can be isolated somewhere on phase space. As we have discussed, mutual irreducibility means that the likelihood ratio ranges over its entire domain, $\mathcal{L} \in [0, \infty)$. If these criteria are satisfied, then the quark distribution of the likelihood ratio observable can be defined from the distributions on samples 1 and 2 as:

$$p_q(\mathcal{L}) = \frac{f_g^{(2)} p_1(\mathcal{L}) - f_g^{(1)} p_2(\mathcal{L})}{f_q^{(1)} - f_q^{(2)}} \ . \tag{6.89}$$

Here, $p_i(\mathcal{L})$ is the distribution of observable $\mathcal{L}$ on sample i, and $f_q^{(i)}$ ($f_g^{(i)}$) is the fraction of quark (gluon) jets in sample i, such that $f_q^{(i)} + f_g^{(i)} = 1$. The corresponding gluon distribution is defined analogously, and the jet fractions $f_q^{(i)}, f_g^{(i)}$ can be extracted through a convex optimization problem [513, 104].

Mutual irreducibility is a necessary property for this inverse problem to be well-defined, and, unfortunately, we technically have not identified a quark versus gluon discriminant that is truly mutual irreducible.[35] However, total hadronic multiplicity only has exponentially small contamination to defining a pure quark sample, so it is expected that it should work rather well for this purpose, which is indeed what is observed in simulated data [105]. A limitation of this technique

---

[35]The existence of a positive-definite probability distribution solution to Eq. (6.89) only requires knowing the values of the reducibility factors or asserting them by fiat. However, if the minimum and maximum values of the likelihood are not 0 and ∞, respectively, then the definition of the topics can be sensitive to the particular approximations one employs, and may change with more information or a better approximation. In topic modeling in natural language processing, perfectly pure samples are referred to as **anchor words** that appear exclusively in a writing sample about a given topic. I thank Eric Metodiev for clarification about this point.

is that quark and gluon topic labels are only defined statistically, and not on an event-by-event basis. So, if we only use multiplicity to extract quark and gluon distributions, only quark and gluon distributions of the multiplicity are returned. There may be some methods to clean up the statistical extraction to be closer to an event-by-event level, by, say, providing the topic modeling with more differential distributions on phase space, but that would also suffer from the curse of dimensionality as the resolved dimensionality of phase space increases.

6.3.2 Are Quark and Gluon Particle Multiplicities Related by KNO Scaling?[36]

In our analysis of quark versus gluon discrimination with hadronic multiplicity, we made the assumption that the fundamental KNO function for both, $\psi(x)$, was the same, and all that differed in their functional form were the average multiplicities, $\langle n_q \rangle < \langle n_g \rangle$. This seems like a reasonable enough assumption, but it also seems reasonable that particle production in quark and gluon jets may be different enough that their KNO functions would, and perhaps should, be different. At any rate, this can in principle be tested, and we can see if this assumption is born out in experimental data.

Unfortunately, there isn't much extant experimental data on which (charged) particle multiplicity distributions on quark and gluon jets are extracted separately. There were a few studies of quark versus gluon jet differences measured at LEP, the electron-positron collider that occupied the current LHC ring until 2001. In these studies, $e^+ e^- \to 3$ jets events were identified, and then the two quark jets and one gluon jet were identified by relative energies, angular distributions, and other correlations that are expected at least at leading-order perturbatively. Early first studies [515, 295, 516, 517, 518] measured quark and gluon jet multiplicity distributions at very low jet energies and in some cases determined the parameter k of the negative binomial distribution of the KNO function. While mean values of multiplicities differed, the best fits to negative binomial KNO functions were consistent on quark and gluon jets, but the quoted uncertainties were extremely large.

More definite and precise results were presented in Ref. [519], using similar techniques for quark and gluon jet extraction, but with significantly smaller uncertainties and at higher energies. In Fig. 27, we have plotted the charged particle multiplicity distributions from this study at left, and then the KNO distributions as determined by rescaling by their respective means. While gluon jets naturally have significantly larger uncertainties, two things are rather clear from these data. First, as expected, the mean multiplicity of gluon jets is larger than that of quark jets, and second, though uncertainties are still large, the mean-rescaled KNO distribution for gluon jets is narrower than that for quark jets. In the context of the negative binomial distribution, the value of the k parameter for gluon jets is a bit larger than that for quark jets, $k_g > k_q$.

This observation is supported by a very recent study on quark versus gluon jets by ATLAS [520]. In this study, ATLAS selected jets from distinct samples that are known to have different compositions of quarks and gluons, and was able to extract purified, individual quark and gluon distributions of various observables. One of those observables was the charged track multiplicity, which is displayed in Fig. 28. At left, we show the plot of the multiplicity on quark and gluon jets, and again, as expected, gluons have a larger mean multiplicity than quarks. When re-scaled by their respective means, we see a similar relationship of the KNO functions for quark and gluon jets as was observed in the data at LEP, but with significantly smaller uncertainties: the gluon KNO

_____

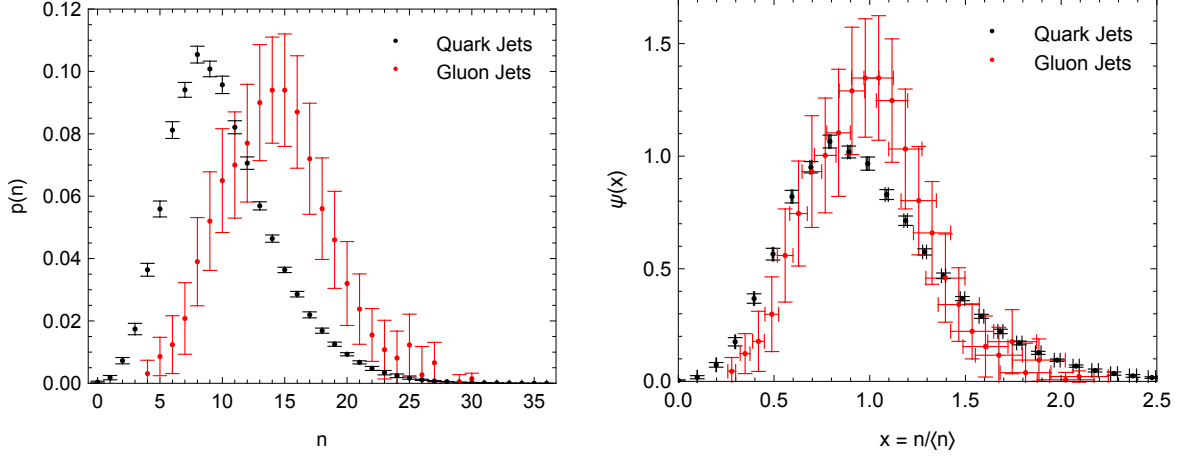[36]Hinchliffe's rule [514] suggests the answer "No", but we will see that it is a bit more subtle.

Figure 27: Plots of the charged particle multiplicity on quark (black) and gluon (red) jets as extracted from $e^+e^- \to 3$ jets events at the OPAL detector at LEP. The energy scale of the jets is about 40-45 GeV. Left: Measured multiplicity distributions. Right: KNO-scaled distributions, in which the distributions have been rescaled by their respective means. Data from Ref. [519].
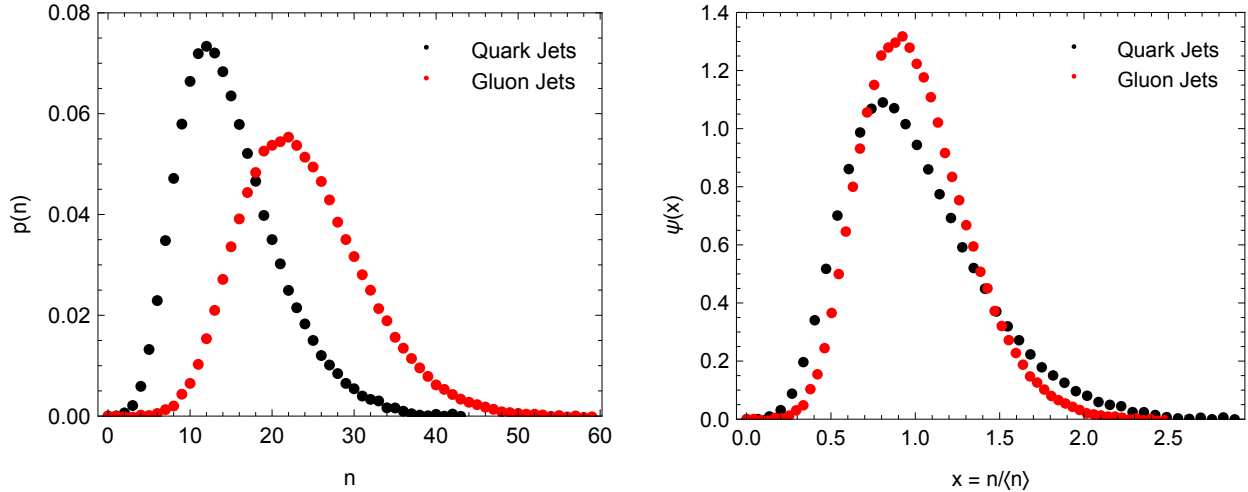

Figure 28: Plots of the charged particle multiplicity (track multiplicity) on quark (black) and gluon (red) jets as extracted from mixed samples with machine learning from jets with transverse momenta of 500-600 GeV in ATLAS events. Uncertainties are at the level of the size of the data points. Left: Measured multiplicity distributions. Right: KNO-scaled distributions, in which the distributions have been rescaled by their respective means. Data extracted from Ref. [520].

function is narrower than that of quarks. Further, the rescaled quark and gluon distributions are very similar between LEP and LHC jets, and each can be fit well by essentially the same value of k in the negative binomial, with $k_q \approx 6$ and $k_g \approx 10$. Further studies will be needed to clearly establish this relationship, and its potential jet energy dependence, as well as provide a theoretical

understanding of its origin, but this is a promising direction for improving discrimination of quark and gluon jets.

Exercises

1. Redo the Lund plane analysis and calculation of the distribution of an angularity τ_α when there is an additional perturbative cut on emissions, $zE\theta > k_\perp$. Is it still true that quark versus gluon discrimination power is independent of angular exponent α? If not, what is the value of α that, say, minimizes the corresponding AUC?

2. Consider the simultaneous measurement of primary Lund plane multiplicity and angularity τ_1. We choose τ_1 so that its contour is parallel to the perturbative cutoff and demand that $\tau_1 > k_\perp/E$. Calculate the joint probability distribution of the angularity τ_1 and the number of emissions in the Lund plane above the scale $k_\perp$ on quark and gluon jets, $p_q(\tau_1, k)$. Show that the likelihood ratio is still monotonic in the multiplicity, and so also measuring the angularity adds no more discrimination information.
   Hint: Don't forget to write the joint distribution with conditional probabilities: $p(\tau_1, k) = p(k|\tau_1)p(\tau_1)$. Where on the Lund plane are the emissions that need to be considered in calculating $p(k|\tau_1)$?

3. From Eq. (6.57), we showed that the AUC for KNO scaling multiplicity takes the form
   $$\mathrm{AUC}(\mathcal{C}) = \int dx\, \psi(x)\, \Psi(\mathcal{C}x)\,, \tag{6.90}$$
   where $0 < \mathcal{C} < 1$ is the ratio of quark to gluon mean multiplicities. Note that this satisfies the differential equation
   $$\frac{d\mathrm{AUC}(\mathcal{C})}{d\mathcal{C}} = \int dx\, x\, \psi(x)\, \psi(\mathcal{C}x)\,, \tag{6.91}$$
   with $\mathrm{AUC}(\mathcal{C} = 1) = 1/2$. Using methods like that in Sec. 6.2.3, provide a robust bound on the integral on the right of Eq. (6.91), as a function of $\mathcal{C}$. Solve the corresponding bounded differential equation for the AUC. How does the resulting bound compare to the explicit value of the AUC with the negative binomial model, Eq. (6.82)?

7 Fourth Example: Quark Flavor Tagging and IRC Unsafe Binary Discrimination

For our final example and lecture, we will pivot to another, distinct, binary jet discrimination problem: classifying jets initiated by different quark flavors. Like general quark versus gluon jet discrimination, this problem has a lot of motivation from every possible direction in collider physics, from pdf extractions, to extractions of Yukawa couplings, to flavor-sensitive new physics. For concreteness, we will restrict our focus to the discrimination of jets initiated by quarks in the first generation, up versus down quark jet discrimination. This problem is of central importance to the goals of the Electron-Ion Collider (EIC) for high-precision extraction of up and down quark pdfs [521].[37]

---

[37]As of the writing of these notes, the EIC is also the only future collider located in the United States that will definitely be constructed and take data.

Another, rather orthogonal, motivation for this problem has come from a rather new corner of collider physics: quantum entanglement. The idea of quantitatively demonstrating that correlations between measurements are due to honest quantum entanglement, and not hidden classical correlations [522, 523], of course dates back to John Bell's thought experiments in the 1960s [524, 525]. Measurements at a collider experiment are classical, all observables commute with one another because only energy is measured, and so this would seem to render any such ability to observe violation of Bell's inequalities impossible. However, the exclusive left-handed nature of the weak interactions can provide a direct and undiluted connection between particle flavor and spin. This has motivated several groups to study flavor correlations in top–anti-top pair production events at the LHC (see, e.g., Refs. [526, 527, 528, 529, 530, 531]), and then correspondingly reinterpreting them through entanglement of the spins of the top quarks at the point of production. These techniques have extremely recently been used by ATLAS and CMS to demonstrably observe quantum entanglement in these systems [532, 533].[38]

These techniques have mostly focused on the case of leptonic top quark decay, for which the charged lepton from subsequent W boson decay has perfect correlation of its momentum with the spin of the initial top quark. This is experimentally very clean, because charged leptons are easily identified and their momenta measured to high precision. However, the top quark pair production decays to visible leptons (electrons or muons) only about 5% of the time, and so one takes a huge statistical hit in restricting to this decay channel. The vast majority of the time, the top quark pairs decay hadronically or semi-leptonically, and so one needs techniques to efficiently analyze these processes to make the most of the extant data. Through the subsequent decay of the W boson, the down-type quark carries perfect correlation of its momentum to the spin of the top, and so an efficient up versus down flavor discriminant could significantly open up the phase space for studying this fundamental of quantum phenomena. Surprisingly rather little work has focused on hadronic top decays (see, e.g., recent Refs. [538, 539, 540]), with perhaps the largest step forward from Ref. [541] that identified the optimal direction of the spin correlation vector from only measuring the total momentum of the jets from top quark decay. If we can construct a powerful up versus down discriminant, perhaps this story will change.

7.1 How Do We Define "Flavor"?

Our goal is to identify the likelihood ratio observable for jets initiated by up versus down quarks. As we did with the quark versus gluon or $H \to b\bar{b}$ versus $g \to b\bar{b}$ problems, we need to first establish what it is we mean when we say "up quark" or "down quark". Naturally, by quantum mechanics, the only observable quantities that can uniquely define a particle are eigenvalues of Hermitian operators. There are but a limited few Hermitian operators whose eigenvalues are actually (potentially) observable in a collider, so let's simply enumerate them for up and down quarks, and then comment on their consequences:

---

[38]Really, these experimental results have only demonstrated strong top quark spin correlations, not unambiguously quantum entanglement because there are still potential loopholes. Further, Bell's unambiguous criteria for entanglement of a pair of particles is expressed as an inequality, but for entanglement of three particles, a boolean test of entanglement is possible [534]. If such a test of entanglement could be extended to collider experimental detectors, there would be no denying the quantum mechanics in your face [535, 536, 537].

Observable	Up Quark	Down Quark
Energy-Momentum	$m_u \approx 2$ MeV	$m_d \approx 4.5$ MeV
Spin	1/2	1/2
Color Charge	C_F	C_F
Electric Charge	$2/3e$	$-1/3e$

There simply isn't much to uniquely define the quarks.[39] Both up and down quarks have exceedingly small masses (values from the PDG [453]) and such a mass is completely irrelevant both because hadron masses are so much greater than individual quark masses, and because any reasonable jet energy is tens, hundreds, or thousands of GeV. Up and down quarks are both fermions, and so have identical spin quantum numbers, so there are no games to play with polarization states, for example, to distinguish them. All quarks carry QCD color in the fundamental representation, with quadratic Casimir C_F, and so the radiation pattern of gluons off of these quarks will be indistinguishable. However, there is one (!!!) quantum number that is distinct and will have observable consequences: electric charge.

Non-zero electric charge is relatively easy to measure in a collider, because high-energy charged particles will emit ionization radiation that is observable and their trajectory will be bent when passing through a magnetic field, with curvature proportional to the electric charge. This would suggest that the net electric charge of a jet,

$$Q_{\text{jet}} = \sum_i Q_i \,, \tag{7.1}$$

where Q_i is the electric charge in units of the fundamental charge e, would be a good discrimination observable for this task. There are, however, at least two problems with interpretation of this raw jet charge. First, every charged particle in the jet contributes with equal strength, and yet a jet can contain a significant amount of very low energy contamination radiation. This contamination can significantly increase the variance of the distribution of such an observable, decreasing discrimination power. Second, there is a disconnect with the electric charge of bare quarks and that of actual observable hadrons. The electric charge of quarks is some fraction of the fundamental electric charge, but the electric charge of hadrons is necessarily an integral multiple of e. So on any jet of interest, it is simply impossible for its observed electric charge to be $1/3$ e, but perhaps in an ensemble, the mean of the jet charge distribution carries information about the quark charge.

The second problem we will address with our formal analysis in the next section, but for the first problem, there is a relatively easy fix. We just weight the contribution to the jet charge by some power of the particles' energy fraction,

$$\mathcal{Q}_\kappa \equiv \sum_i z_i^\kappa \, Q_i \,. \tag{7.2}$$

This definition of "the" jet charge was introduced long ago [547][40] and experimentally studied in great detail [548, 549, 550, 551, 552, 553, 554, 555, 556, 557, 558, 559, 560, 561, 562, 563, 564, 565,

---

[39]I am glossing over another potential source for discrimination. The other light quark, the strange quark, has the same electric charge as the down quark. As such, mesons formed from strange quarks with first-generation quarks, the kaons, differ in their electric charge. So, observing charged kaons in a jet is strongly correlated with up-type jets. Strange jet tagging has seen some interest recently, e.g., Refs. [542, 543, 544, 545, 546], especially applied to (exclusive) Higgs decays.

[40]A quantitative study of jets and the first Monte Carlo event generator was also presented in this paper, written by Feynman and his post-doc Rick Field. Field is emeritus professor at the University of Florida and the brother of actress Sally Field.

566, 567, 568, 568], and whose interest for LHC and jet physics was initiated in Refs. [569, 570] and subsequently measured on jets at the LHC in Refs. [571, 572, 573, 574, 575]. The parameter $\kappa > 0$ so that arbitrarily soft particles do not contribute to the jet charge. This then renders the jet charge infrared safe, but it is still not collinear safe. Consider, for example, the contribution to the jet charge from a gluon that splits to an exactly collinear $u\bar{u}$ pair. The gluon is electrically neutral, and so does not initially contribute to the jet charge. However, after the splitting the $u\bar{u}$ pair modifies the jet charge by

$$\Delta \mathcal{Q}_\kappa = \frac{2}{3}(z_u^\kappa - z_{\bar{u}}^\kappa)\,. \tag{7.3}$$

This has probability 1 to be non-zero, and so the jet charge is irreducibly IRC unsafe. There are no tricks with multi-differential distributions or Sudakov factors that can save us here. To theoretically understand this observable, we need a totally different approach than Feynman diagrams, Lund planes, etc.

7.1.1 Non-Perturbative Assumptions and Consequences

We are at any rate scientists, and starting from nothing has never scared us before. We just need to identify a good and powerful set of assumptions from which to work and which can be tested in experiment. As mentioned previously, the results we derive here will not follow any perturbative analysis in the coupling α_s. Instead, we will use justifiable non-perturbative assumptions that follow from basic observations about the structure of hadronic jets at high energies and their particle content. The assumptions that we exploit were introduced in Ref. [576] and are as follows:

1. Particles (hadrons) in the jet are produced though identical, independent processes.

2. The multiplicity of particles in the jet n is large.

3. The only particles in the jet are the pions: π^+, π^-, and π^0.

4. We assume exact SU(2) isospin symmetry and so quantities measured on any of the pions have identical distributions.

Exclusively from these assumptions, we will be able to quantitatively produce the functional form of the likelihood ratio for up versus down quark discrimination. Let's see how this is done.

We will start with the goal of determining the probability distribution of the jet charge on jets initiated by quark q itself, $p_q(\mathcal{Q}_\kappa)$. Conditioning on the hadronic multiplicity n in the jet, this distribution can be expressed as

$$p_q(\mathcal{Q}_\kappa) = \int dn\, p_q(\mathcal{Q}_\kappa|n)\, p_q(n)\,. \tag{7.4}$$

We can, of course, say something about the KNO scaling of the multiplicity distribution $p(n)$, but we won't worry about that yet. Let's instead focus on the conditioned distribution, $p_q(\mathcal{Q}_\kappa|n)$. The form of the jet charge is a sum over i.i.d. random variables (by the first assumption), and we are considering the large multiplicity $n \to \infty$ limit (by the second assumption). Therefore, the central limit theorem predicts the functional form of this distribution as a Gaussian:

$$p_q(\mathcal{Q}_\kappa|n) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(\mathcal{Q}_\kappa - \mu_q)^2}{2\sigma^2}}\,, \tag{7.5}$$

109

where we must determine the mean μ_q and variance σ^2.

The mean is just the mean of the jet charge on our ensemble of jets with n constituents, where

$$\mu_q = \left\langle \sum_{i=1}^{n} z_i^{\kappa} Q_i \right\rangle = n\langle z^{\kappa}\rangle\langle Q\rangle = Q_q\langle z^{\kappa}\rangle. \tag{7.6}$$

At right, we note that the mean value of the individual electric charges in the jets initiated by a quark q is Q_q/n, where Q_q is the electric charge of the quark q. $\langle z^{\kappa}\rangle$ is the expectation value of the single particle energy fraction raised to power κ. Right now, we have no constraint on what this should be, but we will fix that soon. The variance of the central limit Gaussian is, correspondingly,[41]

$$\sigma^2 = \left\langle \sum_{i=1}^{n} z_i^{2\kappa} Q_i^2 \right\rangle = \frac{2}{3}n\langle z^{2\kappa}\rangle. \tag{7.7}$$

The argument of the sum in the middle follows from our proof of the central limit theorem in Sec. 2.3, and what is especially important here is that each term is proportional to the square of particle i's electric charge. From assumption 3 and 4, because we only measure pions and assume exact isospin symmetry, $Q_i^2 = 0$ or 1 on pions, and 2/3 of the pions are charged. Hence, at right, the mean number of charged pions in each jet in this ensemble is $2/3n$. This variance also needs no subscript decoration because it is universal and identical for all jets, by the assumption of isospin conservation.

We still have to evaluate these tricky moments of the energy fraction distribution, $\langle z^{\kappa}\rangle$ and $\langle z^{2\kappa}\rangle$. These can be calculated through an additional conditional distribution of energy fraction, $p(z|n)$, where, for example,

$$\langle z^{\kappa}\rangle = \int_0^1 dz\, z^{\kappa}\, p(z|n), \tag{7.8}$$

where the energy fraction $z \in [0,1]$. This conditional distribution is of course normalized,

$$1 = \int_0^1 dz\, p(z|n) \tag{7.9}$$

Because all jets in the ensemble have n total particles, the mean of this distribution is

$$\langle z \rangle = \int_0^1 dz\, z\, p(z|n) = \frac{1}{n}, \tag{7.10}$$

as the sum total energy fraction must be 1. However, no other moments of this distribution are known, so we need another approach.

7.1.2 The Moment Expansion

Because the Gaussian distribution of jet charge only needs moments of this energy fraction distribution, it is useful to work with a representation of the distribution expressed exclusively in terms

---

[41]We are being slightly cavalier here. This expression for the variance assumes that the mean jet charge is 0, which is not true in general. However, the variance is a measure of the spread of the distribution about the mean, but is independent of the particular value of the mean, so we are free to set the mean to 0 to evaluate the variance.

of its moments. The question of consistency of the set of moments of a non-negative, integrable distribution is an ancient problem in probability, called the **moment problem**, with known solutions [577, 578, 579]. Formally, any probability distribution $p(x)$ can always be expanded in a series of δ-functions and its derivatives as

$$p(x) = \sum_{k=0}^{\infty} \frac{(-1)^k \langle x^k \rangle}{k!} \delta^{(k)}(x), \tag{7.11}$$

where $\langle x^k \rangle$ is the kth moment.[42] For the energy fraction distribution at hand, $p(z|n)$, we know the mean, so it is much more useful to represent this expansion in terms of central moments $\mu_{z,k}$, where

$$p(z|n) = \delta\left(z - \frac{1}{n}\right) + \sum_{k=2}^{\infty} \frac{(-1)^k \mu_{z,k}}{k!} \delta^{(k)}\left(z - \frac{1}{n}\right). \tag{7.14}$$

Note that there is no term proportional to the first derivative of the δ-function here. Expanding about the mean is expanding about a stable minima, in exactly the same way as expanding the description of a physical pendulum about its static equilibrium. Oscillations about that minimum are first described by the moment of inertia, which is a quadratic moment of the distribution of the mass of the pendulum. The central moments are defined as

$$\mu_{z,k} = \int_0^1 dz \left(z - \frac{1}{n}\right)^k p(z|n), \tag{7.15}$$

and $\mu_{z,2} = \sigma_z^2$ is the variance.[43]

While you are welcome to incorporate higher moments in the description of the likelihood, we will only explicitly consider through the variance, with

$$p(z|n) = \delta\left(z - \frac{1}{n}\right) + \frac{\sigma_z^2}{2} \delta''\left(z - \frac{1}{n}\right) + \cdots, \tag{7.16}$$

suppressing higher moments.[44] This will be sufficient to draw a number of conclusions. Importantly, we note that the variance is strictly non-negative, $\sigma_z^2 \geq 0$. Using this moment expansion, the κ

---

[42]However, it is possible that the moments do not uniquely determine the probability distribution. For example, the distribution

$$p(x) = \frac{e^{-x^{1/4}}}{24} \left(1 + \epsilon \sin(x^{1/4})\right), \tag{7.12}$$

for $x \in [0, \infty)$ and $|\epsilon| < 1$, has moments

$$\int_0^\infty dx \, x^n \, p(x) = \frac{(4n+3)!}{6}, \tag{7.13}$$

independent of ϵ. The moment problem is unique if the higher moments at worst diverge sufficiently slowly [580, 581], with a uniqueness bound of roughly $\langle x^n \rangle \lesssim (2n)!$.

[43]The solution to the moment problem involves constructing the Hankel matrix of all moments and demanding that all possible determinants of minors of the matrix are positive. This constraint has seen use in a number of recent papers as explanation for positivity of Wilson coefficients in an effective field theory, for example Refs. [582, 583, 584, 585, 586, 587, 588, 589, 590, 591].

[44]In general, we expect that the moment expansion is only asymptotic because δ-functions and its derivatives have support at a single point. However, in practice, unless integrating the distribution $p(z|n)$ against a particularly

moment of the energy fraction distribution is then easy to calculate, where

$$\langle z^\kappa \rangle = n^{-\kappa} \left(1 - \frac{\kappa(1-\kappa)}{2} \sigma_z^2 n^2 + \cdots \right) \tag{7.18}$$

Using this result, we can then express the mean and variance of the multiplicity-conditioned jet charge distribution as

$$\mu_q = Q_q n^{-\kappa} \left(1 - \frac{\kappa(1-\kappa)}{2} \sigma_z^2 n^2 + \cdots \right) , \tag{7.19}$$

$$\sigma^2 = \frac{2}{3} n^{1-2\kappa} \left(1 - \kappa(1-2\kappa) \sigma_z^2 n^2 + \cdots \right) . \tag{7.20}$$

In general, the moment expansion of $p(z|n)$ will be an expansion in central moments times a power of multiplicity as $\mu_{z,k} n^k$. The mean of $p(z|n)$ is $\langle z \rangle = 1/n$ and it has support on $z \in [0,1]$ and so general results imply that the largest possible value that the central moment can be is $\mu_{z,k} \sim 1/n$ [592]. In the large n limit, terms in such an expansion would grow unbounded with n and therefore be rather useless. However, this $1/n$ scaling of the central moment $\mu_{z,k}$ corresponds to the distribution

$$p_{\text{worst}}(z|n) = \left(1 - \frac{1}{n} \right) \delta(z) + \frac{1}{n} \delta(1-z) , \tag{7.21}$$

which is highly degenerate and unphysical, with a single particle carrying the entire energy of the jet. A realistic distribution $p(z|n)$ will have most of its support around the mean value, with a tail extending to large values, with strictly 0 probability for $z = 1$. An example of such a distribution would be

$$p(z|n) = (n-1)(1-z)^{n-2} , \tag{7.22}$$

which is normalized and has mean $1/n$. For such a distribution, its kth central moment is

$$\mu_{z,k} = \int_0^1 dz \left(z - \frac{1}{n} \right)^k (n-1)(1-z)^{n-2} \propto n^{-k} , \tag{7.23}$$

so that in the large multiplicity n limit, $\mu_{z,k} n^k$ is order-1 and the moment expansion is sensible. Thus, we have replaced a perturbative expansion in the coupling α_s with a perturbative expansion in central moments.

It's worth saying a bit more about the moment expansion to provide a bit more justification as to where it comes from and why it actually exists. To do this, let's assume we are considering the probability distribution $p(x)$ of a random variable $x \in (-\infty, \infty)$ with mean $\langle x \rangle = 0$. We can, of

---

pathological weight, the moment expansion converges quite rapidly. For example, the mean value of $\log x$ on the uniform distribution on $x \in [0,1]$ is $\langle \log x \rangle = -1$. Using the central moment expansion for the uniform distribution, the first three non-zero terms produce

$$\langle \log x \rangle = \int_0^1 dx \, \log x = -\log 2 - \frac{1}{6} - \frac{1}{20} + \cdots \approx -0.9098 . \tag{7.17}$$
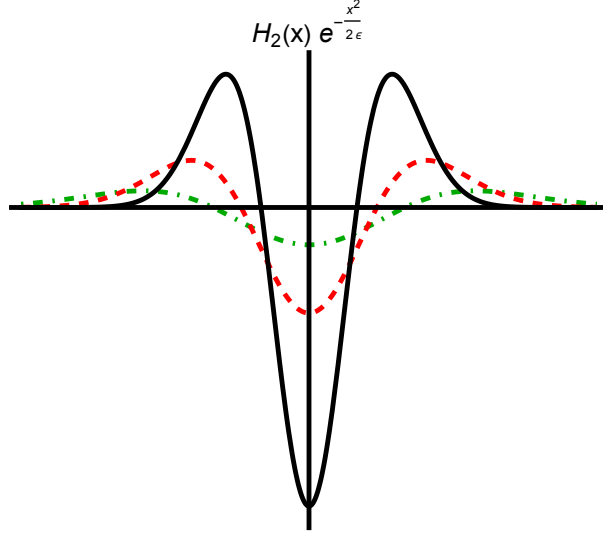
$H_2(\text{x})\, e^{-\frac{x^2}{2\epsilon}}$

Figure 29: Plot of the second Hermite polynomial, as defined in Eq. (7.25), times the Gaussian with variance ϵ. ϵ is decreased by a factor of 2 from green–dot-dashed to red-dashed, and then from red-dashed to black-solid.

course, expand the probability distribution in Hermite polynomials $H_n(x)$ with an overall Gaussian centered at $x = 0$, where

$$p(x) = \frac{1}{\sqrt{2\pi\epsilon}}\, e^{-\frac{x^2}{2\epsilon}} \sum_{i=0}^{\infty} c_i\, H_i(x)\,. \tag{7.24}$$

Here, ϵ is the variance of the Gaussian, c_i are some coefficients, and the Hermite polynomials are defined through a (non-standard) Rodrigues-like formula as

$$H_n(x) = e^{\frac{x^2}{2\epsilon}}\, \frac{d^n}{dx^n} e^{-\frac{x^2}{2\epsilon}}\,. \tag{7.25}$$

The functional form of the second Hermite polynomial with this definition times the Gaussian kernel is plotted in Fig. 29, as the variance ϵ decreases.

With this expansion, note that the Gaussian becomes the δ-function in the limit that $\epsilon \to 0$. With the Rodrigues formula, we can determine the coefficients of the expansion c_i through their relationship to moments. Let's calculate the contribution to moment $\langle x^m \rangle$ from the nth order Hermite polynomial. We have

$$\langle x^m \rangle \supset \frac{c_n}{\sqrt{2\pi\epsilon}} \int_{-\infty}^{\infty} dx\, x^m\, e^{-\frac{x^2}{2\epsilon}} H_n(x) = \frac{c_n}{\sqrt{2\pi\epsilon}} \int_{-\infty}^{\infty} dx\, x^m\, \frac{d^n}{dx^n} e^{-\frac{x^2}{2\epsilon}}\,. \tag{7.26}$$

Now, we can integrate-by-parts to move all the derivatives from the Gaussian onto the power of x. Note that if $n > m$, these derivatives eventually kill x^m, so we only need to consider $m \geq n$. We then have

$$\langle x^m \rangle \supset (-1)^n \frac{m!}{(m-n)!} \frac{c_n}{\sqrt{2\pi\epsilon}} \int_{-\infty}^{\infty} dx\, x^{m-n}\, e^{-\frac{x^2}{2\epsilon}}\,. \tag{7.27}$$

113

We could just evaluate the integral that remains, but it is more enlightening to rescale the integration variable to remove dependence on ϵ. With $x \to \sqrt{\epsilon}x$, we then have

$$\langle x^m \rangle \supset (-1)^n \frac{m!}{(m-n)!} \epsilon^{\frac{m-n}{2}} \frac{c_n}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx\, x^{m-n}\, e^{-\frac{x^2}{2}} \,. \tag{7.28}$$

We already noted that $m \geq n$, and with our δ-function limit in mind, we can now take the $\epsilon \to 0$ limit. If $m > n$, then all such contributions are suppressed by positive powers of ϵ, so only a single term remains: $n = m$. Thus,

$$\langle x^m \rangle = \lim_{\epsilon \to 0} \sum_{n=0}^{m} (-1)^n \frac{m!}{(m-n)!} \epsilon^{\frac{m-n}{2}} \frac{c_n}{\sqrt{2\pi}} \int_{-\infty}^{\infty} dx\, x^{m-n}\, e^{-\frac{x^2}{2}} = m!(-1)^m c_m \,. \tag{7.29}$$

Therefore, the Hermite polynomial expansion of the probability distribution can be expressed in the limit that:

$$p(x) = \lim_{\epsilon \to 0} \frac{1}{\sqrt{2\pi\epsilon}}\, e^{-\frac{x^2}{2\epsilon}} \sum_{i=0}^{\infty} c_i\, H_i(x) = \delta(x) + \sum_{k=2}^{\infty} \frac{(-1)^k \langle x^k \rangle}{k!}\, \delta^{(k)}(x) \,. \tag{7.30}$$

So, while the derivative of δ-function expansion for a probability distribution may seem awkward and ill-defined, it is simply the narrow Gaussian limit of the familiar expansion into orthogonal Hermite poynomials.

7.1.3   Measurements of Discrimination Power and the Likelihood Ratio

With these results for the Gaussian distributions of jet charge conditioned on multiplicity, we can then analyze and quantify its discrimination power. We will start with the distance between the mean values μ_u and μ_d with respect to the square-root of the sum of the variances of the jet charge distributions. As we demonstrated earlier, this quantity is the fundamental discrimination power distance and appears, for example, in the argument of error function in the calculation of the AUC. Using the means and variances we calculated above in the moment expansion of the energy fraction distribution $p(z|n)$, we have

$$\frac{|\mu_u - \mu_d|}{\sqrt{\sigma^2 + \sigma^2}} = \frac{n^{-\kappa} \left(1 - \frac{\kappa(1-\kappa)}{2} \sigma_z^2 n^2 + \cdots \right)}{\sqrt{\frac{4}{3} n^{1-2\kappa} \left(1 - \kappa(1-2\kappa) \sigma_z^2 n^2 + \cdots \right)}} \tag{7.31}$$

$$= \frac{\sqrt{3}}{2\sqrt{n}} \left(1 - \frac{\kappa^2}{2} \sigma_z^2 n^2 + \cdots \right) \,.$$

This expression makes two concrete predictions that can be validated in (simulated) data. First, there is overall dependence on the inverse of the multiplicity, and so the distance between the jet charge distributions increases as multiplicity n decreases. There is some evidence in previous machine learning studies of the jet charge that discrimination improves as multiplicity decreases [593], but this can be directly demonstrated in simulated data by binning jets by multiplicity [576]. Second, κ dependence appears as the coefficient proportional to the variance of the energy fraction distribution, and the discrimination power increases as κ decreases. This property of the jet charge has been directly observed in several prior studies [569, 593, 594, 594, 595, 596], but only recently
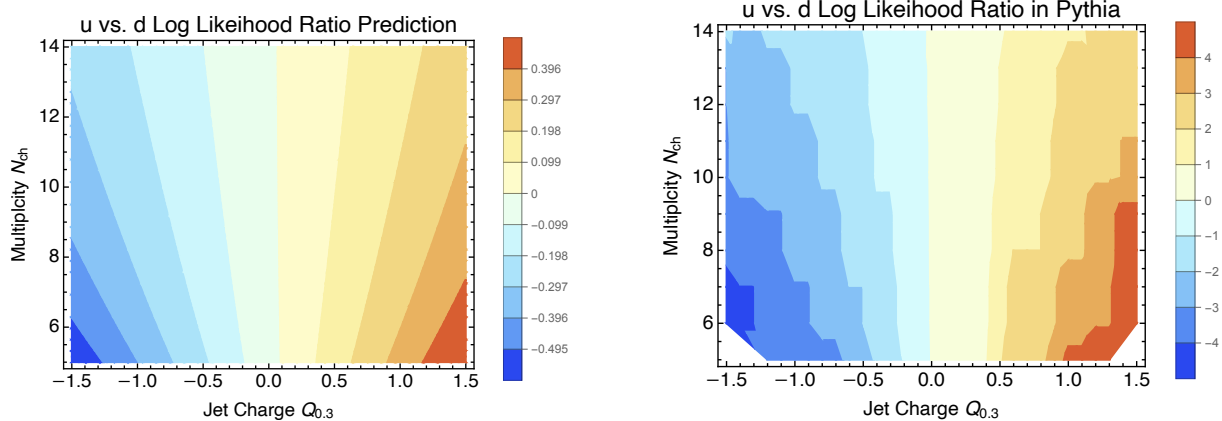
114

Figure 30: Contours of the logarithm of the likelihood ratio on the plane of the charged particle multiplicity N_{ch} and the jet charge $\mathcal{Q}_{\kappa=0.3}$. At left are the contours from the leading terms of Eq. (7.34), and at right are the contours from simulated data. Figure from Ref. [576].

was this simple explanation derived. We note that both of these observations have limits, simply from the assumptions we made. Specifically, if multiplicity n is too small, then the central limit theorem no longer applies, and the Gaussian form of the jet charge distribution breaks down. Similarly, if energy weighting exponent κ is too small, then jet charge loses its infrared safety and additionally necessarily takes only integer values, because in this limit the jet charge reduces to a sum over individual particles' integer charges.

Let's now move to evaluating the likelihood ratio for up and down quark jets on which the jet charge $\mathcal{Q}_{\kappa}$ and the hadronic multiplicity n is measured. From both our assumption of isospin conservation and the quantum numbers of up and down quarks, the multiplicity distributions of hadrons in up and down quark jets will be identical:

$$p_u(n) = p_d(n) \equiv p(n) \,. \tag{7.32}$$

Then, the likelihood ratio as the ratio of the joint probability distributions of the jet charge and multiplicity reduces to the ratio of the jet charge distributions conditioned on the multiplicity:

$$\mathcal{L} = \frac{p_u(\mathcal{Q}_{\kappa}, n)}{p_d(\mathcal{Q}_{\kappa}, n)} = \frac{p_u(\mathcal{Q}_{\kappa}|n)p(n)}{p_d(\mathcal{Q}_{\kappa}|n)p(n)} = \frac{p_u(\mathcal{Q}_{\kappa}|n)}{p_d(\mathcal{Q}_{\kappa}|n)} \,. \tag{7.33}$$

These conditional distributions are Gaussians by the central limit theorem, so it is useful to take the logarithm, which is a monotonic function and so is still the optimal discriminant. The logarithm of the likelihood ratio using the results derived above is then

$$\log \mathcal{L} = \frac{3}{2}n^{-1+\kappa}\mathcal{Q}_{\kappa} - \frac{1}{4n} + \frac{\kappa}{4n}\left(\kappa + 3n^{\kappa}\mathcal{Q}_{\kappa}(1-3\kappa)\right)\sigma_z^2 n^2 + \cdots \,, \tag{7.34}$$

where we have expanded through the variance contribution of the energy fraction distribution.

The rather remarkable thing about this expression is that the likelihood ratio is not monotonic in the jet charge $\mathcal{Q}_{\kappa}$; it is a non-trivial function of both jet charge and multiplicity. Thus, studies that use the jet charge alone for discrimination of jets initiated by up versus down quarks are

115

necessarily suboptimal, and can be improved by additionally measuring particle multiplicity. This prediction, which again follows from the eminently reasonable, basic assumptions of things like the central limit theorem and isospin conservation, can be tested in simulated data. This comparison is shown in Fig. 30. At left is plotted the contours of the log likelihood on the plane of the jet charge $\mathcal{Q}_{\kappa=0.3}$ versus charged particle multiplicity, N_{ch}. By isospin, the charged particle multiplicity is just $2/3$ of the total multiplicity. Here, we have just included the first two terms of Eq. (7.34), ignoring the variance σ_z^2 term for simplicity. Indeed, these contours fan out as multiplicity increases, demonstrating the non-trivial interrelation of these observables. At right on this figure is plotted the log likelihood as extracted from simulated data of jets initiated by up and down quarks. The data set is finite, and so there are some residual binning effects, and further, the scale of the contours is off by a factor (that is beyond the domain of our assumptions), but the shapes of the predicted and simulated contours are strikingly similar.

7.1.4   Up and Down Quark Reducibility Factors

As we have in other situations, the expression for the likelihood itself provides a lower bound estimate on the AUC through the value of its minimum and maximum, the reducibility factors. Naively, we would think that the jet charge in our approximation is Gaussian distributed and therefore can range over all values, $\mathcal{Q}_\kappa \in (-\infty, \infty)$, and so the likelihood ranges over all allowed values, $\mathcal{L} \in [0, \infty)$. However, the jet charge $\mathcal{Q}_\kappa$ is conditioned on the multiplicity of particles in the jet n, and this fixes finite maximum and minimum values of it. This correspondingly will affect the bounds on the likelihood, as we will derive now.[45]

First, note that the expression for the log likelihood of the jet charge and multiplicity, Eq. (7.34), is monotonic in the jet charge, for fixed multiplicity. That is, at fixed multiplicity, the bounds on the likelihood are set by the bounds on the jet charge itself. For fixed multiplicity n, recall that the jet charge is

$$\mathcal{Q}_\kappa = \sum_{i=1}^{n} z_i^\kappa Q_i \,. \tag{7.35}$$

With our assumption that jets only consist of pions, the maximum value of the jet charge is accomplished in a jet full of π^+:

$$\mathcal{Q}_\kappa \leq \sum_{i=1}^{n} z_i^\kappa \,. \tag{7.36}$$

Assuming that $\kappa \leq 1$, the sum over energy fractions is maximized if all are equal, $z_i = 1/n$, and so the jet charge is bounded from above as[46]

$$\mathcal{Q}_\kappa \leq n^{1-\kappa} \,. \tag{7.37}$$

---

[45]We are working within the constraints and assumptions of the central limit theorem, which is only an accurate description of the jet charge distribution in the large n limit, and relatively close to its mean. The bounds of the likelihood are determined by the behavior far in the tails, where the Gaussian approximation is not valid. Nevertheless, we will use these assumptions to derive first quantitative results here as a benchmark for future improved studies. I thank Eric Metodiev for suggesting this analysis.

[46]If $\kappa > 1$, the sum over energy fractions is maximized when a single particle takes all of the energy, so the corresponding bound on the jet charge is $\mathcal{Q}_{\kappa>1} \leq 1$.

The minimum value of the jet charge is just the opposite of this, if the jet is filled with π^-. Thus, the bounds on the value of the jet charge at fixed multiplicity are

$$-n^{1-\kappa} \leq \mathcal{Q}_\kappa \leq n^{1-\kappa}. \tag{7.38}$$

Let's use this to determine the bounds on the log likelihood. With the maximum value of the jet charge, the log likelihood is bounded from above as

$$\log \mathcal{L} \leq \frac{3}{2} - \frac{1}{4n} + \frac{\kappa}{4n}\left(\kappa + 3n(1 - 3\kappa)\right)\sigma_z^2 n^2 + \cdots. \tag{7.39}$$

To maximize this expression over multiplicity n now, we see that we need to take the $n \to \infty$ limit in which this reduces to

$$\log \mathcal{L} \leq \frac{3}{2} + \frac{3}{4}\kappa(1 - 3\kappa)\sigma_z^2 n^2 + \cdots, \tag{7.40}$$

where we assume that $\sigma_z^2 n^2 \sim 1$, as discussed earlier. For the minimum value of the log likelihood, we replace $\mathcal{Q}_\kappa \to -n^{1-\kappa}$ and have

$$\log \mathcal{L} \geq -\frac{3}{2} - \frac{1}{4n} + \frac{\kappa}{4n}\left(\kappa - 3n(1 - 3\kappa)\right)\sigma_z^2 n^2 + \cdots. \tag{7.41}$$

This is minimized for $n \to 0$, for which the likelihood vanishes. This is obviously very far from the large multiplicity assumption, but nevertheless, let's take $n = 0$ here for the expression of the minimum value of the likelihood, but a more honest consideration and consequences of the small multiplicity limit would be fascinating and important for this problem.

Then, the bounds on the log likelihood for up versus down quark jet discrimination with the jet charge and multiplicity are

$$-\infty \leq \log \mathcal{L} \leq \frac{3}{2} + \frac{3}{4}\kappa(1 - 3\kappa)\sigma_z^2 n^2 + \cdots, \tag{7.42}$$

again, assuming $\kappa \leq 1$. These bounds can then be used to establish a robust lower bound on the AUC, from the expression of Eq. (2.30), which we reproduce here:

$$\text{AUC} \geq \frac{1 - 2\mathcal{L}_{\min} + \mathcal{L}_{\max}\mathcal{L}_{\min}}{2(\mathcal{L}_{\max} - \mathcal{L}_{\min})}. \tag{7.43}$$

With sufficiently small κ, the subleading contributions to the bounds of the likelihood are negligible, and so we will estimate the bounds of the likelihood with its leading terms:

$$0 \leq \mathcal{L} \lesssim e^{3/2} \approx 4.48. \tag{7.44}$$

While a pure sample of down quark jets is possible because the minimum likelihood is 0, a pure sample of up quark jets is not because the upper bound is not infinite. The corresponding bound on the AUC is then

$$\text{AUC} \gtrsim \frac{1}{2e^{3/2}} \approx 0.112. \tag{7.45}$$

This will be an interesting baseline for comparison with quantitative calculations of the AUC from the constructed distributions of the jet charge and multiplicity.

7.2 Effects of Multiplicity Distribution

Now that we have the likelihood ratio for up versus down quark discrimination, let's get our hands dirty and actually calculate some explicit discrimination metrics on the various distributions. Here, we will present calculations of the AUCs for both the measurement of the jet charge alone, and when we measure the likelihood ratio of jet charge and multiplicity, as established above. In these calculations, we will further be able to include effects of non-trivial multiplicity distributions and demonstrate what effect a finite width in the KNO distribution for the multiplicity has on discrimination. This will produce additional dependence on the energy weight exponent κ and, combined with earlier results with a finite width of the energy fraction distribution $p(z|n)$, provides a potential explanation of observed κ dependence in discrimination power.

7.2.1 Calculating the Jet Charge AUC

Let's start by just calculating the AUC for up versus down jet discrimination with the jet charge alone. The integrals we need to calculate are

$$\mathrm{AUC}_{\mathcal{Q}_\kappa} = \int d\mathcal{Q}_u \, d\mathcal{Q}_d \, p_u(\mathcal{Q}_u) \, p_d(\mathcal{Q}_d) \, \Theta(\mathcal{Q}_d - \mathcal{Q}_u) \,, \tag{7.46}$$

but there are additional integrals implicit. The jet charge distribution, inclusive over particle multiplicity, is

$$p(\mathcal{Q}_\kappa) = \int dn \, p(\mathcal{Q}_\kappa | n) \, p(n) \,, \tag{7.47}$$

and the conditional distribution takes the familiar Gaussian form. Then, the AUC in all of its glory is

$$\mathrm{AUC}_{\mathcal{Q}_\kappa} = \int d\mathcal{Q}_u \, d\mathcal{Q}_d \, p_u(\mathcal{Q}_u) \, p_d(\mathcal{Q}_d) \, \Theta(\mathcal{Q}_d - \mathcal{Q}_u) \tag{7.48}$$

$$= \int d\mathcal{Q}_u \, d\mathcal{Q}_d \, dn_u \, dn_d \, p_u(\mathcal{Q}_u|n_u) p(n_u) \, p_d(\mathcal{Q}_d|n_d) p(n_d) \, \Theta(\mathcal{Q}_d - \mathcal{Q}_u)$$

$$= \int d\mathcal{Q}_u \, d\mathcal{Q}_d \, dn_u \, dn_d \, \frac{1}{\sqrt{2\pi \frac{2}{3} n_u^{1-2\kappa}}} \frac{1}{\sqrt{2\pi \frac{2}{3} n_d^{1-2\kappa}}} \exp\left[-\frac{\left(\mathcal{Q}_u - \frac{2}{3} n_u^{-\kappa}\right)^2}{\frac{4}{3} n_u^{1-2\kappa}} - \frac{\left(\mathcal{Q}_d + \frac{1}{3} n_d^{-\kappa}\right)^2}{\frac{4}{3} n_d^{1-2\kappa}} \right]$$

$$\times p(n_u) \, p(n_d) \, \Theta(\mathcal{Q}_d - \mathcal{Q}_u) \,.$$

With our goal of establishing the leading dependence on the shape of the multiplicity distribution $p(n)$, we just use the leading terms for the mean and variance of the Gaussians. I encourage you to evaluate the AUC with more general expressions, however!

We can explicitly integrate over the jet charges $\mathcal{Q}_u, \mathcal{Q}_d$ which is made easier with the change of variables

$$x = n_u^\kappa \mathcal{Q}_u \,, \qquad\qquad y = n_d^\kappa \mathcal{Q}_d \,. \tag{7.49}$$

This eliminates much of the κ dependence, but cannot remove it completely:

$$\mathrm{AUC}_{\mathcal{Q}_\kappa} = \int dx \, dy \, dn_u \, dn_d \, \frac{1}{\sqrt{2\pi \frac{2}{3} n_u}} \frac{1}{\sqrt{2\pi \frac{2}{3} n_d}} \exp\left[-\frac{\left(x - \frac{2}{3}\right)^2}{\frac{4}{3} n_u} - \frac{\left(y + \frac{1}{3}\right)^2}{\frac{4}{3} n_d} \right] \tag{7.50}$$

$$\times\, p(n_u)\, p(n_d)\, \Theta(n_d^{-\kappa} y - n_u^{-\kappa} x)$$

$$= \frac{1}{2} - \frac{1}{2} \int dn_u\, dn_d\, p(n_u)\, p(n_d)\, \mathrm{erf}\left(\frac{n_u^\kappa + 2n_d^\kappa}{2\sqrt{3}\sqrt{n_u^{2\kappa} n_d + n_d^{2\kappa} n_u}} \right) .$$

For generic random variables n_u, n_d, the argument of the error function in the remaining integral over multiplicities is not monotonic in κ, so to establish the κ dependence of discrimination power requires actually integrating over the multiplicity distributions. Nevertheless, at this point, we can directly observe that discrimination power of the jet charge is necessarily dependent on the energy scaling exponent κ, which is consistent with our earlier analysis.

We will integrate over the multiplicity distributions shortly, but at this stage it is also useful to stare at some expressions for particularly simple, limiting choices of κ. While $\kappa = 0$ has some problems with interpretation as discussed earlier, its AUC is rather simple:

$$\mathrm{AUC}_{\mathcal{Q}_0} = \frac{1}{2} - \frac{1}{2} \int dn_u\, dn_d\, p(n_u)\, p(n_d)\, \mathrm{erf}\left(\frac{\sqrt{3}}{2\sqrt{n_d + n_u}} \right) . \tag{7.51}$$

In the limit that $\kappa = 0$, all particles contribute equally to the jet charge, independent of energy, and note that here, the argument of the error function is a democratic sum of multiplicities, $n_u + n_d$. In the opposite limit, $\kappa \to \infty$, in which the jet charge is fixed by the hardest particle in the jet alone, its AUC is

$$\mathrm{AUC}_{\mathcal{Q}_\infty} \tag{7.52}$$
$$= \frac{1}{2} - \frac{1}{2} \int dn_u\, dn_d\, p(n_u)\, p(n_d) \left[\Theta(n_u - n_d)\, \mathrm{erf}\left(\frac{1}{2\sqrt{3}\sqrt{n_d}} \right) + \Theta(n_d - n_u)\, \mathrm{erf}\left(\frac{1}{\sqrt{3}\sqrt{n_u}} \right) \right] .$$

Only keeping the leading expression for the mean and variance of the jet charge distributions effectively means that the energy fraction distribution is just a δ-function:

$$p(z|n) = \delta\left(z - \frac{1}{n} \right) . \tag{7.53}$$

Therefore, if the multiplicity of the up quark jet is larger than the down quark jet, $n_u > n_d$, then the hardest particle in the down quark jet is more energetic than that in the up quark jet, and we see that the argument of the error function only depends on the multiplicity n_d in that case.

7.2.2 Calculating the Jet Charge and Multiplicity Likelihood AUC

Let's now move to the calculation of the AUC for jets on which the likelihood ratio of jet charge and multiplicity is measured. By the Neyman-Pearson lemma, we "know" that this must be smaller than measuring the jet charge alone, but seeing how it works in action is rather illuminating and informative. Our first task is to calculate the distribution of the likelihood itself, where

$$p(\mathcal{L}) = \int d\mathcal{Q}_\kappa\, dn\, p(\mathcal{Q}_\kappa|n)\, p(n)\, \delta\left(\mathcal{L} - \frac{3}{2} n^{-1+\kappa} \mathcal{Q}_\kappa + \frac{1}{4n} + \cdots \right) . \tag{7.54}$$

As discussed earlier, this is actually the logarithm of the likelihood ratio (which eliminates extreme exponential suppression), and we will work, as before, to leading-order in the expansion of the

likelihood in the moments of the energy fraction distribution. The AUC of the likelihood can then be expressed as

$$\text{AUC}_{\mathcal{L}} = \int d\mathcal{L}_u \, d\mathcal{L}_d \, p_u(\mathcal{L}_u) \, p_d(\mathcal{L}_d) \, \Theta(\mathcal{L}_d - \mathcal{L}_u) \tag{7.55}$$

$$= \int d\mathcal{Q}_u \, d\mathcal{Q}_d \, dn_u \, dn_d \, \frac{1}{\sqrt{2\pi \frac{2}{3} n_u^{1-2\kappa}}} \frac{1}{\sqrt{2\pi \frac{2}{3} n_d^{1-2\kappa}}} \exp\left[-\frac{\left(\mathcal{Q}_u - \frac{2}{3} n_u^{-\kappa}\right)^2}{\frac{4}{3} n_u^{1-2\kappa}} - \frac{\left(\mathcal{Q}_d + \frac{1}{3} n_d^{-\kappa}\right)^2}{\frac{4}{3} n_d^{1-2\kappa}} \right]$$

$$\times p(n_u) \, p(n_d) \, \Theta\left(\frac{3}{2} n_d^{-1+\kappa} \mathcal{Q}_d - \frac{1}{4n_d} - \frac{3}{2} n_u^{-1+\kappa} \mathcal{Q}_u + \frac{1}{4n_u} \right) .$$

While these integrals still look like a huge mess, we can re-introduce the scaled charge variables

$$x = n_u^\kappa \mathcal{Q}_u \,, \qquad\qquad y = n_d^\kappa \mathcal{Q}_d \,, \tag{7.56}$$

which, in this case of the likelihood, completely eliminates all explicit κ dependence on discrimination:

$$\text{AUC}_{\mathcal{L}} = \int dx \, dy \, dn_u \, dn_d \, \frac{1}{\sqrt{2\pi \frac{2}{3} n_u}} \frac{1}{\sqrt{2\pi \frac{2}{3} n_d}} \exp\left[-\frac{\left(x - \frac{2}{3}\right)^2}{\frac{4}{3} n_u} - \frac{\left(y + \frac{1}{3}\right)^2}{\frac{4}{3} n_d} \right] \tag{7.57}$$

$$\times p(n_u) \, p(n_d) \, \Theta\left(\frac{6y - 1}{n_d} - \frac{6x - 1}{n_u} \right)$$

$$= \frac{1}{2} - \frac{1}{2} \int dn_u \, dn_d \, p(n_u) \, p(n_d) \, \text{erf}\left(\frac{\sqrt{3}}{4} \sqrt{\frac{1}{n_u} + \frac{1}{n_d}} \right) .$$

This is a rather amazing result. It is true that if you only measure the jet charge, then discrimination power is a function of κ, as, as argued earlier, one should choose reasonably small κ for the best performance. However, if you measure jet charge and multiplicity of the jets, then the resulting likelihood ratio has discrimination power performance that is independent of κ! It has been observed that the variance in discrimination power among jet charges with different κ is significantly smaller when the jet charge is but one input to a neural network that also includes all particle momenta (and also, therefore, total multiplicity), e.g., Ref. [593].

7.2.3 Optimality of the Likelihood

Let's now compare these results from the measurement of the jet charge to the measurement of the likelihood ratio to explicitly demonstrate the latter's dominance as a discriminant for up versus down jet classification. The first comparison we will make is of the integrand of the AUC for $\kappa = 0$ jet charge and the likelihood. We first make the following observation: that the geometric mean of two numbers n_u, n_d is always bounded by their arithmetic mean:

$$\sqrt{n_u n_d} \le \frac{n_u + n_d}{2} \,. \tag{7.58}$$

This is equivalent to

$$\sqrt{\frac{n_u n_d}{n_u + n_d}} \le \frac{\sqrt{n_u + n_d}}{2} \,, \tag{7.59}$$

or, on inverting this relationship and sprinkling in a couple of numerical factors,

$$\frac{\sqrt{3}}{2\sqrt{n_u + n_d}} \le \frac{\sqrt{3}}{4}\sqrt{\frac{1}{n_u} + \frac{1}{n_d}}\,. \tag{7.60}$$

This final expression is a direct comparison of the arguments of the error functions in the integrands of the AUCs for $\mathcal{Q}_0$ and the likelihood ratio, Eqs. (7.51) and (7.57). The error function monotonically increases as a function of its argument, so this demonstrates directly that $\text{AUC}_{\mathcal{Q}_0} > \text{AUC}_{\mathcal{L}}$.

As mentioned earlier, this simple comparison of integrands doesn't work for generic values of κ and instead the remaining integrals over the multiplicities n_u, n_d must be evaluated to establish optimal discrimination performance with the likelihood ratio. To do these integrals, we need an expression for the multiplicity distribution $p(n)$ and through our detailed study of KNO scaling, we have such an expression. However, the negative binomial, for example, would produce nasty integrals whose dependence on the relevant parameters would be significantly muddied, so let's again do something extremely simple, but nevertheless robust. As it worked for us before, let's again use the moment expansion, but this time for the multiplicity distribution about its mean, where

$$p(n) = \delta(n - \langle n \rangle) + \frac{\sigma_n^2}{2}\delta''(n - \langle n \rangle) + \cdots, \tag{7.61}$$

and, as before, we will only explicitly work through the variance.

Now, with this simple expression for the multiplicity distribution, we can evaluate the AUCs for jet charge with arbitrary κ and the likelihood. We find

$$\text{AUC}_{\mathcal{Q}_\kappa} = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{\sqrt{3}}{2\sqrt{2}}\frac{1}{\sqrt{\langle n \rangle}}\right) + \frac{\sigma_n^2}{2\langle n \rangle^2}\frac{9 - 84\langle n \rangle + 4\kappa^2 + 48\langle n \rangle(\kappa - 1)^2}{64\sqrt{6\pi}\langle n \rangle^{3/2}}\,e^{-\frac{3}{8\langle n \rangle}} + \cdots, \tag{7.62}$$

for the jet charge and for the likelihood, we find

$$\text{AUC}_{\mathcal{L}} = \frac{1}{2} - \frac{1}{2}\text{erf}\left(\frac{\sqrt{3}}{2\sqrt{2}}\frac{1}{\sqrt{\langle n \rangle}}\right) + \frac{\sigma_n^2}{2\langle n \rangle^2}\frac{9 - 84\langle n \rangle}{64\sqrt{6\pi}\langle n \rangle^{3/2}}\,e^{-\frac{3}{8\langle n \rangle}} + \cdots. \tag{7.63}$$

Here, we have scaled the variance σ_n^2 by the squared mean, as, from KNO scaling, we expect that $\sigma_n^2/\langle n \rangle^2 \sim 1$. The first two terms of both are equal, so we just need to compare the third terms, proportional to the variance of the multiplicity distribution. The κ dependent terms in the jet charge AUC are manifestly positive, and so indeed the AUC for the likelihood is smaller and therefore is a better discriminant than the jet charge alone.

The AUC for the jet charge suggests a rather interesting optimal choice for κ, due to the finite width of the multiplicity distribution. For any reasonable average multiplicity $\langle n \rangle$, the $(\kappa - 1)^2$ term will overwhelmingly dominate the AUC unless $\kappa = 1$, and so it is clear that this is the optimal value to minimize the AUC. Note that this result is not at odds with what we found above, that discrimination power increases as $\kappa \to 0$. This previous result assumed a finite width of the particle energy fraction distribution with fixed multiplicity, while this current result assumes a fixed and equal energy fraction for each particle but finite width of the multiplicity distribution. Actually, combining these results may lead to better insight for other results that suggest that small, but non-zero values of κ are actually optimal for jet charge [569, 593]. That is, these competing effects of finite energy fraction width that prefers $\kappa \to 0$ and finite multiplicity width that prefers $\kappa \to 1$ are both optimized for intermediate κ, $\kappa \sim 0.3 - 0.5$, or so.
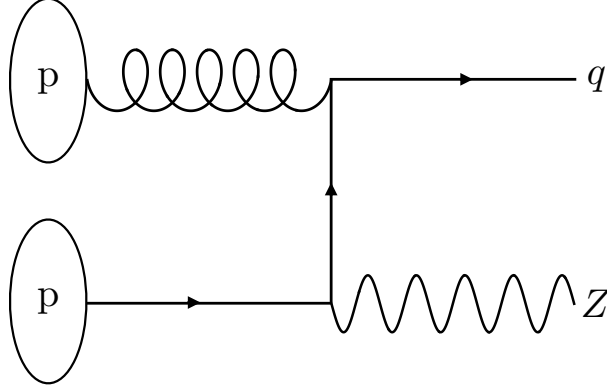
Figure 31: Feynman diagram of $pp \to Z+$ jet events in which a quark parton q is produced at leading order.

7.3 Quark versus Gluon Jet Discrimination One Final Time

As we argued in the previous section, if all you measure are particle momenta, then the hadronic multiplicity itself is an extremely good quark versus gluon discrimination observable, and in the double logarithmic limit, is in fact the likelihood ratio for this problem. With jet charge, however, there is potentially more useful information for this problem because gluons are electrically neutral while quarks are electrically charged, and so, at least on average the jet charge distributions from quark or gluon jets will be displaced from one another. In this section, we will explore how multiplicity and the jet charge can be used in concert to further improve our quark versus gluon jet discriminant.

Before diving into calculations, there are a few potential confusions to address. First, in a fully inclusive quark jet sample, where all flavors of quarks initiate jets with equal probability, the average jet charge would be 0. In your quark jet sample, there would be just as many jets initiated by up quarks as anti-up quarks, and these obviously have opposite electric charge. However, some simple selection criteria can significantly bias the quark jet flavor; for example, by studying the jet in $pp \to Z+$ jet events. The leading-order Feynman diagram for this process is illustrated in Fig. 31. Here, the flavor of the quark jet q is exactly the flavor of the parton pulled out of the proton in collision. The proton is composed of up and down valence quarks, and so dominantly these quark flavors are represented in the jet. If we just naively assume that the fraction of up and down quarks that initiate the jet in these events is described by the relative quark content in the proton, then 2/3 of the time the jet is up flavor, while 1/3 of the time it is down. The average electric charge of the jet is then

$$\langle Q_j \rangle = \frac{2}{3} Q_u + \frac{1}{3} Q_d = \frac{2}{3} \cdot \frac{2}{3} + \frac{1}{3} \left(-\frac{1}{3} \right) = \frac{1}{3} . \tag{7.64}$$

This is non-zero, and so would be distinct from the charge of gluon jets, at least on average.

For our analysis in this section, we will simply assume that some selection criteria have been made such that gluon jets have 0 average electric charge, while quark jets have some non-zero average charge, $\langle \mathcal{Q} \rangle_q$. Other than this and the assumption of KNO scaling of the multiplicity, we will again work within the assumptions that we laid out in Sec. 7.1.1.

122

7.3.1 Jet Charge Distributions on Quarks and Gluons and Discrimination Measures

Our first step to analyzing this problem is to determine the joint probability distributions of jet charge and multiplicity on quark and gluon jets. This can be expressed as the product of the jet charge distribution conditioned on multiplicity and the multiplicity itself as

$$p_q(\mathcal{Q}_\kappa, n) = p_q(\mathcal{Q}_\kappa|n)p_q(n) = \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left[-\frac{(\mathcal{Q}_\kappa - \langle\mathcal{Q}\rangle_q)^2}{2\sigma_q^2}\right] p_q(n)\,, \qquad (7.65)$$

for quarks, and

$$p_g(\mathcal{Q}_\kappa, n) = p_g(\mathcal{Q}_\kappa|n)p_g(n) = \frac{1}{\sqrt{2\pi\sigma_g^2}} \exp\left[-\frac{\mathcal{Q}_\kappa^2}{2\sigma_g^2}\right] p_g(n)\,, \qquad (7.66)$$

for gluons. Here, we have assumed that the widths of the jet charge distributions are not equal for quark and gluon jets, but they can nevertheless be expressed in terms of moments of the corresponding energy fraction distributions, where

$$\sigma_q^2 = \frac{2}{3}n\langle z^{2\kappa}\rangle_q\,, \qquad\qquad\qquad \sigma_g^2 = \frac{2}{3}n\langle z^{2\kappa}\rangle_g\,. \qquad (7.67)$$

Here, the energy fraction moment is defined with respect to the corresponding energy fraction distribution from a jet of flavor f as

$$\langle z^{2\kappa}\rangle_f = \int_0^1 dz\, z^{2\kappa}\, p_f(z|n)\,. \qquad (7.68)$$

The average value of the quark's jet charge can be expressed similarly, where

$$\langle\mathcal{Q}\rangle_q = Q_q\langle z^\kappa\rangle_q\,. \qquad (7.69)$$

As before, we use the moment expansions for the energy fraction distributions through the variance,

$$p_q(z|n) = \delta\left(z - \frac{1}{n}\right) + \frac{\sigma_{q,z}^2}{2}\delta''\left(z - \frac{1}{n}\right) + \cdots\,, \qquad (7.70)$$

$$p_g(z|n) = \delta\left(z - \frac{1}{n}\right) + \frac{\sigma_{g,z}^2}{2}\delta''\left(z - \frac{1}{n}\right) + \cdots\,. \qquad (7.71)$$

We assume that the widths of these distributions are different on quark and gluon jets, $\sigma_{g,z}^2 \neq \sigma_{g,z}^2$, and further, there is evidence that the energy distribution on gluon jets is narrower, $\sigma_{q,z}^2 > \sigma_{g,z}^2$, which can be extracted from fragmentation functions and mean multiplicity measurements of quark and gluon jets from e^+e^- collisions [518, 597, 598]. That is, gluons fragment into hadrons that carry more similar energies than in comparable quark jets.[47]

---

[47]Here, we assume that the average jet charge on quark jets is non-zero. However, one could instead consider the case in which both quark and gluon jets have zero mean jet charge, and only the widths of the jet charge distributions differ. This would produce a different expression for the likelihood ratio, but would still produce improved discrimination performance over multiplicity alone.
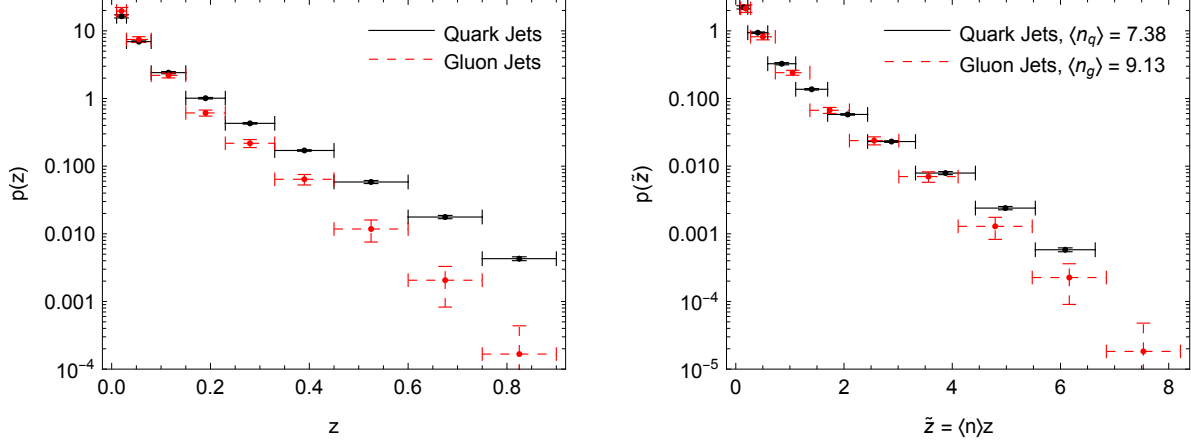
Figure 32: Plots of the energy fraction distribution $p(z)$ from $e^+e^- \to 3$ jets events on quark (black) and gluon (red) jets with energies of about 25 GeV. At left are the normalized energy fraction distributions and at right are those distributions rescaled by the corresponding charged particle multiplicities. Data from Ref. [598] and mean multiplicities from Ref. [518].

Evidence for this is presented in Fig. 32. At left, we plot the energy fraction distributions $p(z)$ as measured on quark and gluon jets in $e^+e^- \to 3$ jets events [598], where the jet energy is about 25 GeV. The gluon jet distribution is clearly narrower than that of quark jets, but we also expect that the particle multiplicity of gluon jets is larger than that of quarks. Because of this fact, the average energy fraction $\langle z \rangle$ inclusive in the number of particles n in the jet is expected to scale inversely with the mean multiplicity:

$$\langle z \rangle = \int dz\, dn\, z\, p(z|n)\, p(n) = \int dn\, \frac{p(n)}{n} = \frac{1}{\langle n \rangle} + \cdots, \tag{7.72}$$

where at right, we have expanded the multiplicity distribution $p(n)$ to lowest order in the moment expansion,

$$p(n) = \delta(n - \langle n \rangle) + \cdots. \tag{7.73}$$

For sensitivity to the variance of the exclusive energy fraction distribution, $p(z|n)$, we rescale the inclusive distribution by the mean multiplicity, $\tilde{z} = \langle n \rangle z$, at right in Fig. 32. While the differences between quarks and gluons are significantly reduced with this rescaling, the gluon jet distribution is still slightly narrower than that of quark jets, suggesting that $\sigma_{q,z}^2 > \sigma_{g,z}^2$. However, much clearer evidence would be provided by simply binning in particle multiplicity, but to my knowledge, such data do not exist.

With these results, the mean values of the conditional jet charge distribution are

$$\mu_q = Q_q n^{-\kappa} \left(1 - \frac{\kappa(1-\kappa)}{2}\, \sigma_{q,z}^2 n^2 + \cdots \right), \qquad \mu_g = 0\,, \tag{7.74}$$

and the variances are

$$\sigma_q^2 = \frac{2}{3} n^{1-2\kappa} \left(1 - \kappa(1-2\kappa)\, \sigma_{q,z}^2 n^2 + \cdots \right), \tag{7.75}$$

124

$$\sigma_g^2 = \frac{2}{3}n^{1-2\kappa}\left(1 - \kappa(1-2\kappa)\,\sigma_{g,z}^2 n^2 + \cdots\right).$$

The distance between means as measured by the sum of the variances of the conditioned jet charge distributions is then

$$\frac{|\mu_q - \mu_g|}{\sqrt{\sigma_q^2 + \sigma_g^2}} = \frac{\sqrt{3}}{2}\frac{|Q_q|}{\sqrt{n}}\left[1 - \frac{\kappa}{4}\left(1 - (1-2\kappa)\frac{\sigma_{g,z}^2}{\sigma_{q,z}^2}\right)\sigma_{q,z}^2 n^2 + \cdots\right]. \tag{7.76}$$

This distance as a measure of the discrimination power of jet charge and multiplicity is rather interesting. The overall scaling with the multiplicity is similar to that for up versus down discrimination, and shows that as multiplicity decreases, this distance between distributions increases, and so too does discrimination power.

The κ dependence is much less obvious, but let's take a derivative of the κ-dependent piece and see what we find. We have

$$\frac{d}{d\kappa}\left[-\kappa\left(1 - (1-2\kappa)\frac{\sigma_{g,z}^2}{\sigma_{q,z}^2}\right)\right] = -1 + (1-4\kappa)\frac{\sigma_{g,z}^2}{\sigma_{q,z}^2}. \tag{7.77}$$

Because we assume that $\sigma_{g,z}^2 < \sigma_{q,z}^2$, this is necessarily negative for all $\kappa > 0$, and so discrimination power is still optimized for small κ, $\kappa \to 0$.

7.3.2 The Likelihood Ratio and Its AUC

While it would be interesting to study the dependence on the quark versus gluon jet AUC as a function of the shape of the energy fraction distributions, we will restrict our study here to the leading scaling dependence. That is, we will just take

$$\langle z^{2\kappa}\rangle_f = n^{-2\kappa} + \cdots, \tag{7.78}$$

and so the functional form of the distributions we work with are

$$p_q(\mathcal{Q}_\kappa, n) = \frac{1}{\sqrt{2\pi\frac{2}{3}n^{1-2\kappa}}}\exp\left[-\frac{(\mathcal{Q}_\kappa - Q_q n^{-\kappa})^2}{\frac{4}{3}n^{1-2\kappa}}\right]p_q(n), \tag{7.79}$$

$$p_g(\mathcal{Q}_\kappa, n) = \frac{1}{\sqrt{2\pi\frac{2}{3}n^{1-2\kappa}}}\exp\left[-\frac{\mathcal{Q}_\kappa^2}{\frac{4}{3}n^{1-2\kappa}}\right]p_g(n). \tag{7.80}$$

With these distributions, we then find that the logarithm of the likelihood ratio is

$$\log\mathcal{L} = \log\frac{p_g(\mathcal{Q}_\kappa, n)}{p_q(\mathcal{Q}_\kappa, n)} = -\frac{3}{2}Q_q n^{-1+\kappa}\mathcal{Q}_\kappa + \frac{3}{4}\frac{Q_q^2}{n} + \cdots + \log\frac{p_g(n)}{p_q(n)}. \tag{7.81}$$

As observed with up versus down quark jet discrimination, this likelihood ratio is monotonic in neither the jet charge nor the multiplicity, and so is necessarily a better discriminant than either. The ellipses hide dependence on the explicit moments of the energy fraction distributions, $p_q(z|n)$ and $p_g(z|n)$.

With these restrictions, the AUC for quark versus gluon discrimination with the likelihood ratio of jet charge and multiplicity is

$$\text{AUC}_{\mathcal{Q}_\kappa, n} = \int dn_q\, dn_g\, d\mathcal{Q}_q\, d\mathcal{Q}_g\, p_q(\mathcal{Q}_q, n_q)\, p_q(\mathcal{Q}_g, n_g)\, \Theta\left(\log \frac{p_g(\mathcal{Q}_q, n_q)}{p_q(\mathcal{Q}_q, n_q)} - \log \frac{p_g(\mathcal{Q}_g, n_g)}{p_q(\mathcal{Q}_g, n_g)}\right) \quad (7.82)$$

$$= \int dn_q\, dn_g\, d\mathcal{Q}_q\, d\mathcal{Q}_g\, \frac{1}{2\pi \frac{2}{3}\sqrt{n_q^{1-2\kappa} n_g^{1-2\kappa}}}\, \exp\left[-\frac{(\mathcal{Q}_q - Q_q n_q^{-\kappa})^2}{\frac{4}{3}n_q^{1-2\kappa}} - \frac{\mathcal{Q}_g^2}{\frac{4}{3}n_g^{1-2\kappa}}\right]\, p_q(n_q)\, p_g(n_g)$$

$$\times\, \Theta\left(-\frac{3}{2}Q_q n_q^{-1+\kappa}\mathcal{Q}_q + \frac{3}{4}\frac{Q_q^2}{n_q} + \frac{3}{2}Q_q n_g^{-1+\kappa}\mathcal{Q}_g - \frac{3}{4}\frac{Q_q^2}{n_g} + \log\frac{p_g(n_q)}{p_q(n_q)} - \log\frac{p_g(n_g)}{p_q(n_g)}\right)$$

$$= \int dn_q\, dn_g\, p_q(n_q)\, p_g(n_g)\left[\frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{4n_g n_q \Delta_{\min} - 3(n_q + n_g)Q_q^2}{4\sqrt{3}\sqrt{n_q n_g(n_q + n_g)}|Q_q|}\right)\right].$$

For compactness, we have introduced

$$\Delta_{\min} \equiv \log\frac{p_g(n_q)}{p_q(n_q)} - \log\frac{p_g(n_g)}{p_q(n_g)}. \quad (7.83)$$

We would like to compare this to the situation in which we only measure the multiplicity on the jets, which would have a corresponding AUC of

$$\text{AUC}_n = \int dn_q\, dn_g\, p_q(n_q)\, p_g(n_g)\, \Theta(n_q - n_g). \quad (7.84)$$

Note that to the order of approximations we work, the quark and gluon jet charge distributions have the same width, and only their mean differs by the net charge Q_q. So, if the charge $Q_q = 0$, the quark and gluon jet charge distributions would be identical, and no discrimination information could be gained from measuring the jet charge. Thus, for comparison to just measuring the multiplicity, it makes sense to expand the expression for the AUC of Eq. (7.82) in the value of the net charge Q_q itself, assuming that $|Q_q| \ll 1$, which is expected as discussed earlier. For the most direct comparison of these two expressions for the AUCs, we will need a singular expansion of the error function, in terms of Heaviside Θ-functions and its derivatives (which are Dirac δ-functions and its derivatives). However, from this lecture, we are now professionals at this, and this shouldn't scare us one bit.

Our goal is then to expand the error function as

$$\frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{4n_g n_q \Delta_{\min} - 3(n_q + n_g)Q_q^2}{4\sqrt{3}\sqrt{n_q n_g(n_q + n_g)}|Q_q|}\right) = \Theta(n_q - n_g) + c_1|Q_q|^{a_1}\delta(n_q - n_g) + \cdots, \quad (7.85)$$

where terms with higher powers of Q_q and more derivatives of the δ-function are encoded in the ellipses. We will just work to calculate the coefficient c_1 and power a_1 here. We first need to validate the lowest-order term. To do this, we can simplify the argument of the error function in the $Q_q \to 0$ limit as

$$\lim_{Q_q \to 0}\left[\frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{4n_g n_q \Delta_{\min} - 3(n_q + n_g)Q_q^2}{4\sqrt{3}\sqrt{n_q n_g(n_q + n_g)}|Q_q|}\right)\right] = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{n_q - n_g}{|Q_q|}\right). \quad (7.86)$$

126

Note that $\Delta_{\min}$ is monotonic in $n_q - n_g$ and all other factors of multiplicity in the error function are positive, and so can safely be ignored to this order because they do not affect the ordering of the argument. Now, in the limit that $Q_q \to 0$, the error function rapidly switches from positive to negative values about $n_q - n_g = 0$, and, with the vertical displacement and scaling, exactly corresponds to the step function:

$$\lim_{Q_q \to 0} \left[\frac{1}{2} + \frac{1}{2} \mathrm{erf} \left(\frac{4 n_g n_q \Delta_{\min} - 3(n_q + n_g) Q_q^2}{4\sqrt{3} \sqrt{n_q n_g (n_q + n_g)} |Q_q|} \right) \right] = \lim_{Q_q \to 0} \left[\frac{1}{2} + \frac{1}{2} \mathrm{erf} \left(\frac{n_q - n_g}{|Q_q|} \right) \right] \qquad (7.87)$$

$$= \Theta(n_q - n_g) \,.$$

For the next term, proportional to the δ-function, we take the derivative with respect to $|Q_q|$ in the error function to produce

$$\frac{d}{d|Q_q|} \left[\frac{1}{2} + \frac{1}{2} \mathrm{erf} \left(\frac{4 n_g n_q \Delta_{\min} - 3(n_q + n_g) Q_q^2}{4\sqrt{3} \sqrt{n_q n_g (n_q + n_g)} |Q_q|} \right) \right] \qquad (7.88)$$

$$= -\frac{4 n_g n_q \Delta_{\min} + 3(n_q + n_g) Q_q^2}{4\sqrt{\pi} \sqrt{3} \sqrt{n_q n_g (n_q + n_g)} Q_q^2} \exp \left[-\frac{(4 n_g n_q \Delta_{\min} - 3(n_q + n_g) Q_q^2)^2}{48 n_q n_g (n_q + n_g) Q_q^2} \right] \,.$$

We will first approximate this derivative as

$$-\frac{4 n_g n_q \Delta_{\min} + 3(n_q + n_g) Q_q^2}{4\sqrt{\pi} \sqrt{3} \sqrt{n_q n_g (n_q + n_g)} Q_q^2} \exp \left[-\frac{(4 n_g n_q \Delta_{\min} - 3(n_q + n_g) Q_q^2)^2}{48 n_q n_g (n_q + n_g) Q_q^2} \right] = a_1 \tilde{c}_1 |Q_q|^{a_1 - 1} \, \delta(4 n_g n_q \Delta_{\min}) \,, \qquad (7.89)$$

for some coefficient $\tilde{c}_1$ and we have taken the derivative of the coefficient of the δ-function term, as well. To determine this, we simply integrate both sides over all values of $x \equiv 4 n_g n_q \Delta_{\min}$ and find

$$a_1 \tilde{c}_1 |Q_q|^{a_1 - 1} = -\int_{-\infty}^{\infty} dx \, \frac{x + 3(n_q + n_g) Q_q^2}{4\sqrt{\pi} \sqrt{3} \sqrt{n_q n_g (n_q + n_g)} Q_q^2} \exp \left[-\frac{(x - 3(n_q + n_g) Q_q^2)^2}{48 n_q n_g (n_q + n_g) Q_q^2} \right] \qquad (7.90)$$

$$= -6(n_q + n_g) |Q_q| \,.$$

Matching powers of $|Q_q|$, we see that $a_1 = 2$ and we can expand the error function in powers of the average quark jet charge as

$$\frac{1}{2} + \frac{1}{2} \mathrm{erf} \left(\frac{4 n_g n_q \Delta_{\min} - 3(n_q + n_g) Q_q^2}{4\sqrt{3} \sqrt{n_q n_g (n_q + n_g)} |Q_q|} \right) = \Theta(n_q - n_g) - 3(n_q + n_g) Q_q^2 \, \delta\left(4 n_g n_q \Delta_{\min}\right) + \cdots$$

$$= \Theta(n_q - n_g) - \frac{3}{4} \frac{n_q + n_g}{n_q n_g} Q_q^2 \, \delta \left(\log \frac{p_g(n_q)}{p_q(n_q)} - \log \frac{p_g(n_g)}{p_q(n_g)} \right) + \cdots \,. \qquad (7.91)$$

With this result, this explicitly demonstrates that the AUC when jet charge and multiplicity are measured is smaller than just that for multiplicity:

$$\mathrm{AUC}_{\mathcal{Q}_\kappa, n} - \mathrm{AUC}_n = -\frac{3}{4} Q_q^2 \int dn_q \, dn_g \, p_q(n_q) \, p_g(n_g) \, \frac{n_q + n_g}{n_q n_g} \, \delta \left(\log \frac{p_g(n_q)}{p_q(n_q)} - \log \frac{p_g(n_g)}{p_q(n_g)} \right) + \cdots$$

$$< 0 \,. \qquad (7.92)$$

This result nicely ties together our entire analysis on quark versus gluon jet discrimination. The most striking differences between quarks and gluons are their QCD color charges (which control the number of particles produced in the jets) and their electric charges (which control the average net electric charge of the jets). The observables that are most sensitive to these differences in quantum numbers are the total hadronic multiplicity and the jet charge, and their simultaneous measurement on jets improves discrimination performance over using either individually. We can, of course, continue to refine and improve our picture and definition of what "quark" or "gluon" jets are to squeeze more discrimination power out, but direct sensitivity to color and electric charges is both extremely robust (ultimately sensitive to particle quantum numbers) and, as we have shown, easy to analyze (data space is only two dimensional).

While we will stop here in these lectures, I encourage you to expand the expressions for the AUCs to higher orders in the moment expansions, or to use the explicit negative binomial representation of the KNO function for the multiplicity distribution to quantitatively evaluate the improvement in discrimination performance.

Exercises

1. In our analysis of up versus down quark jet discrimination, we left the value of explicit moments of the energy fraction distribution $p(z|n)$ implicit, to stress the generality of our results. However, it is enlightening to do some calculations within specific models to gain intuition, which we explore here.

   (a) Use the expression for the "simplest" reasonable energy fraction distribution from Eq. (7.22), and redo the analyses of this lecture. Determine exact results from this model, with no moment expansions.

   (b) In the large multiplicity $n \to \infty$ limit, Eq. (7.22) exponentiates which suggests another form of the expansion of this energy fraction distribution as

   $$p(z|n) = n \, e^{-nz} \left[1 + \sum_{i=2}^{\infty} c_i \, L_i(nz) \right] . \tag{7.93}$$

   Here, $L_i(x)$ is a Laguerre polynomial defined as

   $$L_i(x) = \frac{e^x}{i!} \frac{d^i}{dx^i} \left(e^{-x} x^i \right) , \tag{7.94}$$

   and normalized as

   $$\delta_{ij} = \int_0^{\infty} dx \, e^{-x} \, L_i(x) \, L_j(x) . \tag{7.95}$$

   Redo the analyses of this lecture with this Laguerre expansion through the first few orders. What is the value of c_2 if the variance saturates the bound set by the distribution of Eq. (7.21)?

2. Evaluate the AUC for jet charge, Eq. (7.50), and the likelihood ratio, Eq. (7.57), with the negative binomial KNO function for the multiplicity distribution, Eq. (6.79). How do the values of the AUCs depend on the negative binomial parameter k?

3. We mentioned that gluon jets are expected to have smaller dispersion in the energy fraction distribution of their constituent particles than quark jets with fixed particle multiplicity. One interesting observable to consider for quark versus gluon jet discrimination is pTD [599, 600], which is exactly this dispersion:

$$\text{pTD} = \sum_i z_i^2 \,. \tag{7.96}$$

We can, more generally, consider observables that are just the sum of energy fractions raised to the power κ [601]:

$$\lambda^\kappa = \sum_i z_i^\kappa \,. \tag{7.97}$$

Such an observable is just like the jet charge, but without sensitivity to particle electric charge. pTD is λ^2 while $\lambda^1 = 1$ by energy conservation, and $\lambda^0 = n$, the hadronic multiplicity of the jet, but for other values of κ, this is a distinct observable.

Use this observable for quark versus gluon jet discrimination. What value of κ provides the best discrimination power? Does simultaneous measurement of λ^κ and multiplicity n improve discrimination performance beyond that achieved by only measuring multiplicity?

8 Summary and Conclusions

Binary discrimination problems in collider particle physics and especially jet substructure are ubiquitous and a fundamental component of the data inverse problem: given events from collisions at the LHC, what are the physics responsible? This problem has taken on new breadth and depth in this machine learning era, but beyond and beneath all the noise of the parameters and layers and epochs and architecture of some instantiation of a machine learning algorithm are guiding principles, like the Neyman-Pearson lemma or the central limit theorem, that can be used and exploited to establish a principled prediction of discrimination performance. We have used these results to study numerous classification problems in these lectures, reviewing established results in the literature and deriving some novel results.

While we have especially taken time to analyze and re-analyze the problem of discrimination of jets initiated by either quarks or gluons, we also studied dominant Higgs boson decays, optimizing general resonance searches, and up versus down quark jet discrimination. Each of these problems has seen significant, detailed study in the literature. Other than some passing references, however, perhaps the most glaring omission of discrimination problems is that of identification of boosted hadronic decays of top quarks versus massive jets initiated by light QCD partons. Its omission was neither an oversight nor a criticism of its relevance (the HEP ML Living Review alone lists at least 30 references that just study top quark identification), but rather due to its significantly more complicated decay phase space. At leading order, the top quark decays to three partons, and even further enforcing subsequent on-shell decay of the W-boson, phase space for top decay is three dimensional. Further, sensitivity to the flow of color in the decay products of the top only starts with another soft emission, which increases phase space by another three dimensions. We had enough difficulty with two dimensional phase spaces, but perhaps there are simplifications or tricks like those introduced in these lectures that render the problem tractable on a normal sized sheet of paper.

I look forward to observing further progress in both establishing improved performance and deeper understanding of discrimination problems from students of this school or from readers of these lectures.

Acknowledgements

9 References

[1] A. J. Larkoski, I. Moult, and B. Nachman, *Jet Substructure at the Large Hadron Collider: A Review of Recent Advances in Theory and Machine Learning*, Phys. Rept. **841** (2020) 1–63, [`arXiv:1709.04464`].

[2] R. Kogler et al., *Jet Substructure at the Large Hadron Collider: Experimental Review*, Rev. Mod. Phys. **91** (2019), no. 4 045003, [`arXiv:1803.06991`].

[3] D. Guest, K. Cranmer, and D. Whiteson, *Deep Learning and its Application to LHC Physics*, Ann. Rev. Nucl. Part. Sci. **68** (2018) 161–181, [`arXiv:1806.11484`].

[4] K. Albertsson et al., *Machine Learning in High Energy Physics Community White Paper*, J. Phys. Conf. Ser. **1085** (2018), no. 2 022008, [`arXiv:1807.02876`].

[5] A. Radovic, M. Williams, D. Rousseau, M. Kagan, D. Bonacorsi, A. Himmel, A. Aurisano, K. Terao, and T. Wongjirad, *Machine learning at the energy and intensity frontiers of particle physics*, Nature **560** (2018), no. 7716 41–48.

[6] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, *Machine learning and the physical sciences*, Rev. Mod. Phys. **91** (2019), no. 4 045002, [`arXiv:1903.10563`].

[7] D. Bourilkov, *Machine and Deep Learning Applications in Particle Physics*, Int. J. Mod. Phys. A **34** (2020), no. 35 1930019, [`arXiv:1912.08245`].

[8] M. D. Schwartz, *Modern Machine Learning and Particle Physics*, `arXiv:2103.12226`.

[9] G. Karagiorgi, G. Kasieczka, S. Kravitz, B. Nachman, and D. Shih, *Machine Learning in the Search for New Fundamental Physics*, `arXiv:2112.03769`.

[10] A. Boehnlein et al., *Colloquium: Machine learning in nuclear physics*, Rev. Mod. Phys. **94** (2022), no. 3 031003, [`arXiv:2112.02309`].

[11] P. Shanahan et al., *Snowmass 2021 Computational Frontier CompF03 Topical Group Report: Machine Learning*, `arXiv:2209.07559`.

[12] T. Plehn, A. Butter, B. Dillon, T. Heimel, C. Krause, and R. Winterhalder, *Modern Machine Learning for LHC Physicists*, `arXiv:2211.01421`.

[13] B. Nachman et al., *Jets and Jet Substructure at Future Colliders*, Front. in Phys. **10** (2022) 897719, [`arXiv:2203.07462`].

[14] G. DeZoort, P. W. Battaglia, C. Biscarat, and J.-R. Vlimant, *Graph neural networks at the Large Hadron Collider*, Nature Rev. Phys. **5** (2023), no. 5 281–303.

[15] K. Zhou, L. Wang, L.-G. Pang, and S. Shi, *Exploring QCD matter in extreme conditions with Machine Learning*, Prog. Part. Nucl. Phys. **135** (2024) 104084, [`arXiv:2303.15136`].

[16] V. Belis, P. Odagiu, and T. K. Aarrestad, *Machine learning for anomaly detection in particle physics*, Rev. Phys. **12** (2024) 100091, [`arXiv:2312.14190`].

[17] S. Mondal and L. Mastrolorenzo, *Machine Learning in High Energy Physics: A review of heavy-flavor jet tagging at the LHC*, `arXiv:2404.01071`.

[18] M. Feickert and B. Nachman, *A Living Review of Machine Learning for Particle Physics*, `arXiv:2102.02770`.

[19] N. Arkani-Hamed, G. F. Giudice, G. L. Kane, M. J. Strassler, and H. L. Verlinde, *3rd lhc olympic workshop*, 2006.
`https://www.kitp.ucsb.edu/activities/lhco-c06`.

[20] C. Adam-Bourdarios, G. Cowan, C. Germain, I. Guyond, B. Kégl, and D. Rousseau, *Higgs boson machine learning challenge*, 2014.
`https://kaggle.com/competitions/higgs-boson`.

[21] T. Blake, M.-O. Bettler, M. Chrząszcz, F. Dettori, A. Ustyuzhanin, and T. Likhomanenko, *Flavours of physics: Finding $\tau \to \mu\mu\mu$*, 2015.
`https://kaggle.com/competitions/flavours-of-physics`.

[22] D. Rousseau, S. Amrouche, L. Basara, P. Calafiura, V. Estrade, S. Farrell, C. Germain, V. Vava Gligorov, T. Golling, H. Gray, I. Guyon, M. Hushchyn, V. Innocente, M. Kiehn, E. Moyse, A. Salzburger, A. Ustyuzhanin, J.-R. Vlimant, and Y. Yilmaz, *Trackml particle tracking challenge*, 2018.
`https://kaggle.com/competitions/trackml-particle-identification`.

[23] T. Aarrestad et al., *The Dark Machines Anomaly Score Challenge: Benchmark Data and Model Independent Event Classification for the Large Hadron Collider*, SciPost Phys. **12** (2022), no. 1 043, [`arXiv:2105.14027`].

[24] G. Kasieczka et al., *The LHC Olympics 2020 a community challenge for anomaly detection in high energy physics*, Rept. Prog. Phys. **84** (2021), no. 12 124201, [`arXiv:2101.08320`].

[25] "Boost 2024."
`https://agenda.infn.it/event/37093/`.

[26] "Ml4jets 2024."
`https://indico.cern.ch/event/1386125/`.

[27] "Hammer & nails."
`https://indico.cern.ch/event/1202995/`.

[28] "The synergy of scientific and machine learning modelling (syns & ml) workshop."
`https://icml.cc/virtual/2023/workshop/21476`.

[29] "Ai for science: from theory to practice."
`https://neurips.cc/virtual/2023/workshop/66548`.

[30] E. D. Bloom, D. H. Coward, H. DeStaebler, J. Drees, G. Miller, L. W. Mo, R. E. Taylor, M. Breidenbach, J. I. Friedman, G. C. Hartmann, and H. W. Kendall, *High-energy inelastic $e - p$ scattering at 6° and 10°*, Phys. Rev. Lett. **23** (Oct, 1969) 930–934.

[31] M. Breidenbach, J. I. Friedman, H. W. Kendall, E. D. Bloom, D. H. Coward, H. DeStaebler, J. Drees, L. W. Mo, and R. E. Taylor, *Observed behavior of highly inelastic electron-proton scattering*, Phys. Rev. Lett. **23** (Oct, 1969) 935–939.

[32] G. Hanson et al., *Evidence for Jet Structure in Hadron Production by e+ e- Annihilation*, Phys. Rev. Lett. **35** (1975) 1609–1612.

[33] **TASSO** Collaboration, R. Brandelik et al., *Evidence for Planar Events in e+ e- Annihilation at High-Energies*, Phys. Lett. B **86** (1979) 243–249.

[34] D. P. Barber et al., *Discovery of Three Jet Events and a Test of Quantum Chromodynamics at PETRA Energies*, Phys. Rev. Lett. **43** (1979) 830.

[35] **PLUTO** Collaboration, C. Berger et al., *Evidence for Gluon Bremsstrahlung in e+ e- Annihilations at High-Energies*, Phys. Lett. B **86** (1979) 418–425.

[36] **JADE** Collaboration, W. Bartel et al., *Observation of Planar Three Jet Events in e+ e- Annihilation and Evidence for Gluon Bremsstrahlung*, Phys. Lett. B **91** (1980) 142–147.

[37] H. P. Nilles and K. H. Streng, *Quark - Gluon Separation in Three Jet Events*, Phys. Rev. D **23** (1981) 1944.

[38] G. Rochester and C. Butler, *The new unstable cosmic-ray particles*, Reports on Progress in Physics **16** (1953), no. 1 364.

[39] G. F. Sterman, *Partons, factorization and resummation, TASI 95*, in *Theoretical Advanced Study Institute in Elementary Particle Physics (TASI 95): QCD and Beyond*, pp. 327–408, 6, 1995. `hep-ph/9606312`.

[40] G. P. Salam, *Towards Jetography*, Eur. Phys. J. C **67** (2010) 637–686, [`arXiv:0906.1833`].

[41] A. Abdesselam et al., *Boosted Objects: A Probe of Beyond the Standard Model Physics*, Eur. Phys. J. C **71** (2011) 1661, [`arXiv:1012.5412`].

[42] T. Plehn and M. Spannowsky, *Top Tagging*, J. Phys. G **39** (2012) 083001, [`arXiv:1112.4441`].

[43] A. Altheimer et al., *Jet Substructure at the Tevatron and LHC: New results, new tools, new benchmarks*, J. Phys. G **39** (2012) 063001, [`arXiv:1201.0008`].

[44] J. Shelton, *Jet Substructure*, in *Theoretical Advanced Study Institute in Elementary Particle Physics: Searching for New Physics at Small and Large Scales*, pp. 303–340, 2013. `arXiv:1302.0260`.

[45] A. Altheimer et al., *Boosted Objects and Jet Substructure at the LHC. Report of BOOST2012, held at IFIC Valencia, 23rd-27th of July 2012*, Eur. Phys. J. C **74** (2014), no. 3 2792, [`arXiv:1311.2708`].

[46] D. Adams et al., *Towards an Understanding of the Correlations in Jet Substructure*, Eur. Phys. J. C **75** (2015), no. 9 409, [`arXiv:1504.00679`].

[47] M. Cacciari, *Phenomenological and theoretical developments in jet physics at the LHC*, Int. J. Mod. Phys. A **30** (2015), no. 31 1546001, [`arXiv:1509.02272`].

[48] S. Marzani, G. Soyez, and M. Spannowsky, *Looking inside jets: an introduction to jet substructure and boosted-object phenomenology*, vol. 958. Springer, 2019.

[49] R. K. Ellis, W. J. Stirling, and B. R. Webber, *QCD and collider physics*, vol. 8. Cambridge University Press, 2, 2011.

[50] A. J. Larkoski, *An Unorthodox Introduction to QCD*, `arXiv:1709.06195`.

[51] A. J. Larkoski, *Another Unorthodox Introduction to QCD and now Machine Learning*, `arXiv:2008.09673`.

[52] A. J. Larkoski, *Jet Physics from the Ground Up*, `arXiv:2112.15122`.

[53] A. J. Larkoski, *Elementary Particle Physics: An Intuitive Introduction*. Cambridge University Press, 6, 2019.

[54] A. J. Larkoski, *Quantum Mechanics*. Cambridge University Press, 2022.

[55] J. Kaplan, "Notes on Contemporary Machine Learning for Physicists." `https://sites.krieger.jhu.edu/jared-kaplan/files/2019/04/ContemporaryMLforPhysicists.pdf`.

[56] D. A. Roberts, S. Yaida, and B. Hanin, *The Principles of Deep Learning Theory*, `arXiv:2106.10165`.

[57] M. R. Douglas, *Large Language Models*, `arXiv:2307.05782`.

[58] F. Rosenblatt, *The perceptron: a probabilistic model for information storage and organization in the brain.*, *Psychological review* **65** (1958), no. 6 386.

[59] A. G. Ivakhnenko, V. G. Lapa, et al., *Cybernetic predicting devices*, .

[60] S. Amari, *A theory of adaptive pattern classifiers*, *IEEE Transactions on Electronic Computers* (1967), no. 3 299–307.

[61] S. Linnainmaa, *The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors.* PhD thesis, Master's Thesis (in Finnish), Univ. Helsinki, 1970.

[62] K. Fukushima, *Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position*, *Biological cybernetics* **36** (1980), no. 4 193–202.

[63] W. Lenz, *Beitrag zum verständnis der magnetischen erscheinungen in festen körpern*, *Z. Phys.* **21** (1920) 613–615.

[64] E. Ising, *Contribution to the Theory of Ferromagnetism*, *Z. Phys.* **31** (1925) 253–258.

[65] S.-I. Amari, *Learning patterns and pattern sequences by self-organizing nets of threshold elements*, *IEEE Transactions on computers* **100** (1972), no. 11 1197–1206.

[66] W. A. Little, *The existence of persistent states in the brain*, *Mathematical biosciences* **19** (1974), no. 1-2 101–120.

[67] J. J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities.*, *Proceedings of the national academy of sciences* **79** (1982), no. 8 2554–2558.

[68] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*, *Advances in neural information processing systems* **27** (2014).

[69] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, *Attention is all you need*, *Advances in neural information processing systems* **30** (2017).

[70] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark, *Kan: Kolmogorov-arnold networks*, *arXiv preprint arXiv:2404.19756* (2024).

[71] A. Kolmogorov, *On the representation of continuous functions of several variables by superpositions of continuous functions of lesser variable count*, in *Dokl. Akad. Nauk SSSR*, vol. 108, 1956.

[72] V. Arnold, *On representation of any continuous function of three variables as a sum of functions of two variables or less*, in *Dokl. Akad. Nauk SSSR*, vol. 114, p. 4, 1957.

[73] G. Lorentz, *Metric entropy, widths, and superpositions of functions*, *The American Mathematical Monthly* **69** (1962), no. 6 469–485.

[74] D. A. Sprecher, *On the structure of continuous functions of several variables*, *Transactions of the American Mathematical Society* **115** (1965) 340–355.

[75] J. Braun and M. Griebel, *On a constructive proof of kolmogorov's superposition theorem*, *Constructive approximation* **30** (2009) 653–675.

[76] F. Girosi and T. Poggio, *Representation properties of networks: Kolmogorov's theorem is irrelevant*, *Neural Computation* **1** (1989), no. 4 465–469.

[77] V. Kůrková, *Kolmogorov's theorem is relevant*, *Neural computation* **3** (1991), no. 4 617–622.

[78] J.-N. Lin and R. Unbehauen, *On the realization of a kolmogorov network*, *Neural Computation* **5** (1993), no. 1 18–20.

[79] D. A. Sprecher, *A numerical implementation of kolmogorov's superpositions*, *Neural networks* **9** (1996), no. 5 765–772.

[80] D. A. Sprecher, *A numerical implementation of kolmogorov's superpositions ii*, *Neural networks* **10** (1997), no. 3 447–457.

[81] V. Maiorov and A. Pinkus, *Lower bounds for approximation by mlp neural networks*, *Neurocomputing* **25** (1999), no. 1-3 81–91.

[82] M. Köppen, *On the training of a kolmogorov network*, in *Artificial Neural Networks–ICANN 2002: International Conference Madrid, Spain, August 28–30, 2002 Proceedings 12*, pp. 474–479, Springer, 2002.

[83] J. Schmidt-Hieber, *The kolmogorov–arnold representation theorem revisited*, *Neural networks* **137** (2021) 119–126.

[84] V. Dhiman, "Kan: Kolmogorov–arnold networks: A review." https://vikasdhiman.info/reviews/KAN_a_review.pdf.

[85] H. Robbins and S. Monro, *A stochastic approximation method*, *The annals of mathematical statistics* (1951) 400–407.

[86] J. Kiefer and J. Wolfowitz, *Stochastic estimation of the maximum of a regression function*, *The Annals of Mathematical Statistics* (1952) 462–466.

[87] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning internal representations by error propagation, parallel distributed processing, explorations in the microstructure of cognition, ed. de rumelhart and j. mcclelland. vol. 1. 1986*, Biometrika **71** (1986) 599–607.

[88] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, `arXiv:1412.6980`.

[89] G. Cybenko, *Approximation by superpositions of a sigmoidal function*, Mathematics of control, signals and systems **2** (1989), no. 4 303–314.

[90] K. Hornik, *Approximation capabilities of multilayer feedforward networks*, Neural networks **4** (1991), no. 2 251–257.

[91] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, *Multilayer feedforward networks with a nonpolynomial activation function can approximate any function*, Neural networks **6** (1993), no. 6 861–867.

[92] K. Fukushima, *Visual feature extraction by a multilayered network of analog threshold elements*, IEEE Transactions on Systems Science and Cybernetics **5** (1969), no. 4 322–333.

[93] D. Yarotsky, *Error bounds for approximations with deep relu networks*, Neural Networks **94** (2017) 103–114.

[94] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, *The expressive power of neural networks: A view from the width*, Advances in neural information processing systems **30** (2017).

[95] B. Hanin and M. Sellke, *Approximating continuous functions by relu nets of minimal width*, `arXiv:1710.11278`.

[96] P. Kidger and T. Lyons, *Universal approximation with deep narrow networks*, in Conference on learning theory, pp. 2306–2327, PMLR, 2020.

[97] S. Park, C. Yun, J. Lee, and J. Shin, *Minimum width for universal approximation*, `arXiv:2006.08859`.

[98] Z. Shen, H. Yang, and S. Zhang, *Optimal approximation rate of relu networks in terms of width and depth*, Journal de Mathématiques Pures et Appliquées **157** (2022) 101–135.

[99] J. Neyman and E. S. Pearson, *On the Problem of the Most Efficient Tests of Statistical Hypotheses*, Phil. Trans. Roy. Soc. Lond. A **231** (1933), no. 694-706 289–337.

[100] P. Woodward, *Probability and Information Theory: With Applications to Radar*. No. v. 3 in International series of monographs on electronics and instrumentation. McGraw-Hill, 1953.

[101] W. Peterson, T. Birdsall, and W. Fox, *The theory of signal detectability*, Transactions of the IRE Professional Group on Information Theory **4** (1954), no. 4 171–212.

[102] C. E. Shannon, *A mathematical theory of communication*, The Bell system technical journal **27** (1948), no. 3 379–423.

[103] A. J. Larkoski and E. M. Metodiev, *A Theory of Quark vs. Gluon Discrimination*, JHEP **10** (2019) 014, [`arXiv:1906.01639`].

[104] J. Katz-Samuels, G. Blanchard, and C. Scott, *Decontamination of mutual contamination models*, Journal of machine learning research **20** (2019), no. 41.

[105] E. M. Metodiev and J. Thaler, *Jet Topics: Disentangling Quarks and Gluons at Colliders*, *Phys. Rev. Lett.* **120** (2018), no. 24 241602, [`arXiv:1802.00008`].

[106] P. T. Komiske, E. M. Metodiev, and J. Thaler, *An operational definition of quark and gluon jets*, *JHEP* **11** (2018) 059, [`arXiv:1809.01140`].

[107] R. A. Fisher, *On an absolute criterion for fitting frequency curves*, *Messenger of mathematics* **41** (1912) 155–156.

[108] S. S. Wilks, *The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses*, *The Annals of Mathematical Statistics* **9** (1938), no. 1 60 – 62.

[109] S. Kullback and R. A. Leibler, *On information and sufficiency*, *The annals of mathematical statistics* **22** (1951), no. 1 79–86.

[110] T. Bayes, *Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s*, *Philosophical transactions of the Royal Society of London* (1763), no. 53 370–418.

[111] A. De Moivre, *Approximatio ad summum terminorum binomii seriem expansi*, *Proceedings of the Royal Society* (1733).

[112] A. De Moivre, *Doctrine of chance, Millar, London, Reprinted by Chelsea Publishers, New York* (1738) 235–43.

[113] P. S. de Laplace, *Théorie analytique des probabilités*, vol. 7. Courcier, 1820.

[114] A. M. Lyapunov, *Sur un théorème du calcul des probabilités*, *Comptes rendus hebdomadaires des séances de l'Académie des Sciences de Paris* **132** (1901) 126–128.

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

[116] R. M. Neal and R. M. Neal, *Priors for infinite networks*, *Bayesian learning for neural networks* (1996) 29–53.

[117] C. K. Williams, *Computation with infinite neural networks*, *Neural Computation* **10** (1998), no. 5 1203–1216.

[118] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein, *Deep neural networks as gaussian processes*, *arXiv preprint arXiv:1711.00165* (2017).

[119] P. A. M. Dirac, *Quantum theory of emission and absorption of radiation*, *Proc. Roy. Soc. Lond. A* **114** (1927) 243.

[120] E. Fermi, *Nuclear physics: a course given by Enrico Fermi at the University of Chicago*. University of Chicago Press, 1950.

[121] S. Fubini and G. Furlan, *Renormalization effects for partially conserved currents*, *Physics Physique Fizika* **1** (1965), no. 4 229–247.

[122] S. Weinberg, *Dynamics at infinite momentum*, *Phys. Rev.* **150** (1966) 1313–1318.

[123] L. Susskind, *Model of selfinduced strong interactions*, *Phys. Rev.* **165** (1968) 1535–1546.

[124] K. Bardakci and M. B. Halpern, *Theories at infinite momentum*, *Phys. Rev.* **176** (1968) 1686–1699.

[125] S.-J. Chang and S.-K. Ma, *Feynman rules and quantum electrodynamics at infinite momentum*, Phys. Rev. **180** (1969) 1506–1513.

[126] J. B. Kogut and D. E. Soper, *Quantum Electrodynamics in the Infinite Momentum Frame*, Phys. Rev. D **1** (1970) 2901–2913.

[127] P. Cox and T. Melia, *Independently Parameterised Momenta Variables and Monte Carlo IR Subtraction*, JHEP **12** (2018) 038, [`arXiv:1809.09325`].

[128] B. Henning and T. Melia, *Conformal-helicity duality \& the Hilbert space of free CFTs*, `arXiv:1902.06747`.

[129] B. Henning and T. Melia, *Constructing effective field theories via their harmonics*, Phys. Rev. D **100** (2019), no. 1 016015, [`arXiv:1902.06754`].

[130] A. J. Larkoski and T. Melia, *Covariantizing phase space*, Phys. Rev. D **102** (2020), no. 9 094014, [`arXiv:2008.06508`].

[131] T. Cai, J. Cheng, N. Craig, G. Koszegi, and A. J. Larkoski, *The Phase Space Distance Between Collider Events*, `arXiv:2405.16698`.

[132] P. Srivastava and G. Sudarshan, *Multiple production of pions in nuclear collisions*, Phys. Rev. **110** (1958) 765–766.

[133] E. C. G. Sudarshan and R. e. Marshak, *Chirality invariance and the universal Fermi interaction*, Phys. Rev. **109** (1958) 1860–1860.

[134] R. P. Feynman and M. Gell-Mann, *Theory of Fermi interaction*, Phys. Rev. **109** (1958) 193–198.

[135] E. C. G. Sudarshan, *Equivalence of semiclassical and quantum mechanical descriptions of statistical light beams*, Phys. Rev. Lett. **10** (1963) 277–279.

[136] R. J. Glauber, *Coherent and incoherent states of the radiation field*, Phys. Rev. **131** (1963) 2766–2788.

[137] C. Olah, *Neural networks, manifolds, and topology*, 2014. `https://colah.github.io/posts/2014-03-NN-Manifolds-Topology/`.

[138] E. O. Korman, *Autoencoding topology*, arXiv preprint arXiv:1803.00156 (2018).

[139] M. Moor, M. Horn, B. Rieck, and K. Borgwardt, *Topological autoencoders*, in *International conference on machine learning*, pp. 7045–7054, PMLR, 2020.

[140] M. Hajij and K. Istvan, *A topological framework for deep learning*, arXiv preprint arXiv:2008.13697 (2020).

[141] J. Batson, C. G. Haaf, Y. Kahn, and D. A. Roberts, *Topological Obstructions to Autoencoding*, JHEP **04** (2021) 280, [`arXiv:2102.08380`].

[142] M. Beneke and V. A. Smirnov, *Asymptotic expansion of Feynman integrals near threshold*, Nucl. Phys. B **522** (1998) 321–344, [`hep-ph/9711391`].

[143] V. A. Smirnov and E. R. Rakhmetov, *The Strategy of regions for asymptotic expansion of two loop vertex Feynman diagrams*, Theor. Math. Phys. **120** (1999) 870–875, [`hep-ph/9812529`].

[144] V. A. Smirnov, *Problems of the strategy of regions*, Phys. Lett. B **465** (1999) 226–234, [`hep-ph/9907471`].

[145] V. A. Smirnov, *Applied asymptotic expansions in momenta and masses, Springer Tracts Mod. Phys.* **177** (2002) 1–262.

[146] C. W. Bauer, S. Fleming, and M. E. Luke, *Summing Sudakov logarithms in $B \to X_s\gamma$ in effective field theory.*, *Phys. Rev. D* **63** (2000) 014006, [hep-ph/0005275].

[147] C. W. Bauer, S. Fleming, D. Pirjol, and I. W. Stewart, *An Effective field theory for collinear and soft gluons: Heavy to light decays, Phys. Rev. D* **63** (2001) 114020, [hep-ph/0011336].

[148] C. W. Bauer and I. W. Stewart, *Invariant operators in collinear effective theory, Phys. Lett. B* **516** (2001) 134–142, [hep-ph/0107001].

[149] C. W. Bauer, D. Pirjol, and I. W. Stewart, *Soft collinear factorization in effective field theory, Phys. Rev. D* **65** (2002) 054022, [hep-ph/0109045].

[150] C. W. Bauer and I. W. Stewart, "The Soft-Collinear Effective Theory." http://www2.lns.mit.edu/~iains/scetnotes_EFTx_Oct_2014.pdf.

[151] T. Becher, A. Broggio, and A. Ferroglia, *Introduction to Soft-Collinear Effective Theory*, vol. 896.
Springer, 2015.

[152] Y. L. Dokshitzer, *Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics.*, *Sov. Phys. JETP* **46** (1977) 641–653.

[153] V. N. Gribov and L. N. Lipatov, *Deep inelastic e p scattering in perturbation theory, Sov. J. Nucl. Phys.* **15** (1972) 438–450.

[154] V. N. Gribov and L. N. Lipatov, *e+ e- pair annihilation and deep inelastic e p scattering in perturbation theory, Sov. J. Nucl. Phys.* **15** (1972) 675–684.

[155] L. N. Lipatov, *The parton model and perturbation theory, Yad. Fiz.* **20** (1974) 181–198.

[156] G. Altarelli and G. Parisi, *Asymptotic Freedom in Parton Language, Nucl. Phys. B* **126** (1977) 298–318.

[157] J. M. Campbell and E. W. N. Glover, *Double unresolved approximations to multiparton scattering amplitudes, Nucl. Phys. B* **527** (1998) 264–288, [hep-ph/9710255].

[158] S. Catani and M. Grazzini, *Infrared factorization of tree level QCD amplitudes at the next-to-next-to-leading order and beyond, Nucl. Phys. B* **570** (2000) 287–325, [hep-ph/9908523].

[159] Z. Bern, V. Del Duca, and C. R. Schmidt, *The Infrared behavior of one loop gluon amplitudes at next-to-next-to-leading order, Phys. Lett. B* **445** (1998) 168–177, [hep-ph/9810409].

[160] Z. Bern, V. Del Duca, W. B. Kilgore, and C. R. Schmidt, *The infrared behavior of one loop QCD amplitudes at next-to-next-to leading order, Phys. Rev. D* **60** (1999) 116001, [hep-ph/9903516].

[161] D. A. Kosower and P. Uwer, *One loop splitting amplitudes in gauge theory, Nucl. Phys. B* **563** (1999) 477–505, [hep-ph/9903515].

[162] S. D. Badger and E. W. N. Glover, *Two loop splitting functions in QCD, JHEP* **07** (2004) 040, [hep-ph/0405236].

[163] Z. Bern, L. J. Dixon, and D. A. Kosower, *Two-loop g —> gg splitting amplitudes in QCD*, *JHEP* **08** (2004) 012, [`hep-ph/0404293`].

[164] S. Catani, D. de Florian, and G. Rodrigo, *The Triple collinear limit of one loop QCD amplitudes*, *Phys. Lett. B* **586** (2004) 323–331, [`hep-ph/0312067`].

[165] S. Badger, F. Buciuni, and T. Peraro, *One-loop triple collinear splitting amplitudes in QCD*, *JHEP* **09** (2015) 188, [`arXiv:1507.05070`].

[166] V. Del Duca, C. Duhr, R. Haindl, A. Lazopoulos, and M. Michel, *Tree-level splitting amplitudes for a quark into four collinear partons*, *JHEP* **02** (2020) 189, [`arXiv:1912.06425`].

[167] V. Del Duca, C. Duhr, R. Haindl, A. Lazopoulos, and M. Michel, *Tree-level splitting amplitudes for a gluon into four collinear partons*, *JHEP* **10** (2020) 093, [`arXiv:2007.05345`].

[168] M. Czakon and S. Sapeta, *Complete collection of one-loop triple-collinear splitting operators for dimensionally-regulated QCD*, *JHEP* **07** (2022) 052, [`arXiv:2204.11801`].

[169] F. E. Low, *Bremsstrahlung of very low-energy quanta in elementary particle collisions*, *Phys. Rev.* **110** (1958) 974–977.

[170] S. Weinberg, *Infrared photons and gravitons*, *Phys. Rev.* **140** (1965) B516–B524.

[171] T. H. Burnett and N. M. Kroll, *Extension of the low soft photon theorem*, *Phys. Rev. Lett.* **20** (1968) 86.

[172] A. Bassetto, M. Ciafaloni, and G. Marchesini, *Jet Structure and Infrared Sensitive Quantities in Perturbative QCD*, *Phys. Rept.* **100** (1983) 201–272.

[173] G. 't Hooft, *A Planar Diagram Theory for Strong Interactions*, *Nucl. Phys. B* **72** (1974) 461.

[174] G. 't Hooft, *A Two-Dimensional Model for Mesons*, *Nucl. Phys. B* **75** (1974) 461–470.

[175] P. Cvitanovic, *Group theory: birdtracks, Lie's, and exceptional groups*. Princeton University Press, 2008.

[176] F. A. Berends and W. Giele, *The Six Gluon Process as an Example of Weyl-Van Der Waerden Spinor Calculus*, *Nucl. Phys. B* **294** (1987) 700–732.

[177] M. L. Mangano, S. J. Parke, and Z. Xu, *Duality and Multi - Gluon Scattering*, *Nucl. Phys. B* **298** (1988) 653–672.

[178] M. L. Mangano, *The Color Structure of Gluon Emission*, *Nucl. Phys. B* **309** (1988) 461–475.

[179] Z. Bern and D. A. Kosower, *Color decomposition of one loop amplitudes in gauge theories*, *Nucl. Phys. B* **362** (1991) 389–448.

[180] L. J. Dixon, *Calculating scattering amplitudes efficiently*, in *Theoretical Advanced Study Institute in Elementary Particle Physics (TASI 95): QCD and Beyond*, pp. 539–584, 1, 1996. `hep-ph/9601359`.

[181] D. J. Gross and R. Jackiw, *Low-Energy Theorem for Graviton Scattering*, *Phys. Rev.* **166** (1968) 1287–1292.

[182] R. Jackiw, *Low-Energy Theorems for Massless Bosons: Photons and Gravitons*, *Phys. Rev.* **168** (1968) 1623–1633.

[183] C. D. White, *Factorization Properties of Soft Graviton Amplitudes*, *JHEP* **05** (2011) 060, [arXiv:1103.2981].

[184] F. Cachazo and A. Strominger, *Evidence for a New Soft Graviton Theorem*, arXiv:1404.4091.

[185] A. J. Larkoski, *Conformal Invariance of the Subleading Soft Theorem in Gauge Theory*, *Phys. Rev. D* **90** (2014), no. 8 087701, [arXiv:1405.2346].

[186] E. Casali, *Soft sub-leading divergences in Yang-Mills amplitudes*, *JHEP* **08** (2014) 077, [arXiv:1404.5551].

[187] J. Broedel, M. de Leeuw, J. Plefka, and M. Rosso, *Constraining subleading soft gluon and graviton theorems*, *Phys. Rev. D* **90** (2014), no. 6 065024, [arXiv:1406.6574].

[188] Z. Bern, S. Davies, P. Di Vecchia, and J. Nohle, *Low-Energy Behavior of Gluons and Gravitons from Gauge Invariance*, *Phys. Rev. D* **90** (2014), no. 8 084035, [arXiv:1406.6987].

[189] Z. Bern, S. Davies, and J. Nohle, *On Loop Corrections to Subleading Soft Behavior of Gluons and Gravitons*, *Phys. Rev. D* **90** (2014), no. 8 085015, [arXiv:1405.1015].

[190] A. J. Larkoski, D. Neill, and I. W. Stewart, *Soft Theorems from Effective Field Theory*, *JHEP* **06** (2015) 077, [arXiv:1412.3108].

[191] F. Bloch and A. Nordsieck, *Note on the Radiation Field of the electron*, *Phys. Rev.* **52** (1937) 54–59.

[192] T. Kinoshita, *Mass singularities of Feynman amplitudes*, *J. Math. Phys.* **3** (1962) 650–677.

[193] T. D. Lee and M. Nauenberg, *Degenerate Systems and Mass Singularities*, *Phys. Rev.* **133** (1964) B1549–B1562.

[194] G. F. Sterman and S. Weinberg, *Jets from Quantum Chromodynamics*, *Phys. Rev. Lett.* **39** (1977) 1436.

[195] Y.-P. Yao, *On the Infrared Problem in Nonabelian Gauge Theory*, *Phys. Rev. Lett.* **36** (1976) 653.

[196] T. Appelquist, J. Carazzone, H. Kluberg-Stern, and M. Roth, *Infrared Finiteness in Yang-Mills Theories*, *Phys. Rev. Lett.* **36** (1976) 768. [Erratum: Phys.Rev.Lett. 36, 1161 (1976)].

[197] L. Tyburski, *Cancellation of Infrared Divergences in Nonabelian Gauge Theories*, *Phys. Rev. Lett.* **37** (1976) 319–321.

[198] E. C. Poggio and H. R. Quinn, *The Infrared Behavior of Zero-Mass Green's Functions and the Absence of Quark Confinement in Perturbation Theory*, *Phys. Rev. D* **14** (1976) 578.

[199] G. F. Sterman, *Kinoshita's Theorem in Yang-Mills Theories*, *Phys. Rev. D* **14** (1976) 2123–2125.

[200] F. G. Krausz, *Mass Singularities in Yang-Mills Theories*, *Phys. Lett. B* **66** (1977) 251–254.

[201] N. A. Sveshnikov and F. V. Tkachov, *Jets and quantum field theory*, *Phys. Lett. B* **382** (1996) 403–408, [hep-ph/9512370].

[202] F. V. Tkachov, *Measuring multi - jet structure of hadronic energy flow or What is a jet?*, *Int. J. Mod. Phys. A* **12** (1997) 5411–5529, [hep-ph/9601308].

[203] P. S. Cherzor and N. A. Sveshnikov, *Jet observables and energy momentum tensor*, in *12th International Workshop on High-Energy Physics and Quantum Field Theory (QFTHEP 97)*, pp. 402–407, 9, 1997.
`hep-ph/9710349`.

[204] F. V. Tkachov, *A Theory of jet definition*, *Int. J. Mod. Phys. A* **17** (2002) 2783–2884, [`hep-ph/9901444`].

[205] P. T. Komiske, E. M. Metodiev, and J. Thaler, *Energy flow polynomials: A complete linear basis for jet substructure*, *JHEP* **04** (2018) 013, [`arXiv:1712.07124`].

[206] S. Catani, Y. L. Dokshitzer, M. Olsson, G. Turnock, and B. R. Webber, *New clustering algorithm for multi - jet cross-sections in e+ e- annihilation*, *Phys. Lett. B* **269** (1991) 432–438.

[207] S. D. Ellis and D. E. Soper, *Successive combination jet algorithm for hadron collisions*, *Phys. Rev. D* **48** (1993) 3160–3166, [`hep-ph/9305266`].

[208] S. Catani, Y. L. Dokshitzer, M. H. Seymour, and B. R. Webber, *Longitudinally invariant K_t clustering algorithms for hadron hadron collisions*, *Nucl. Phys. B* **406** (1993) 187–224.

[209] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, *Better jet clustering algorithms*, *JHEP* **08** (1997) 001, [`hep-ph/9707323`].

[210] M. Wobisch and T. Wengler, *Hadronization corrections to jet cross-sections in deep inelastic scattering*, in *Workshop on Monte Carlo Generators for HERA Physics (Plenary Starting Meeting)*, pp. 270–279, 4, 1998.
`hep-ph/9907280`.

[211] M. Cacciari, G. P. Salam, and G. Soyez, *The anti-k_t jet clustering algorithm*, *JHEP* **04** (2008) 063, [`arXiv:0802.1189`].

[212] **JADE** Collaboration, W. Bartel et al., *Experimental Studies on Multi-Jet Production in e+ e- Annihilation at PETRA Energies*, *Z. Phys. C* **33** (1986) 23.

[213] **JADE** Collaboration, S. Bethke et al., *Experimental Investigation of the Energy Dependence of the Strong Coupling Strength*, *Phys. Lett. B* **213** (1988) 235–241.

[214] N. Brown and W. J. Stirling, *Jet cross-sections at leading double logarithm in e+ e- annihilation*, *Phys. Lett. B* **252** (1990) 657–662.

[215] N. Brown and W. J. Stirling, *Finding jets and summing soft gluons: A New algorithm*, *Z. Phys. C* **53** (1992) 629–636.

[216] S. Bethke, Z. Kunszt, D. E. Soper, and W. J. Stirling, *New jet cluster algorithms: Next-to-leading order QCD and hadronization corrections*, *Nucl. Phys. B* **370** (1992) 310–334. [Erratum: Nucl.Phys.B 523, 681–681 (1998)].

[217] **UA1** Collaboration, G. Arnison et al., *Observation of Jets in High Transverse Energy Events at the CERN Proton - anti-Proton Collider*, *Phys. Lett. B* **123** (1983) 115–122.

[218] J. E. Huth et al., *Toward a standardization of jet definitions*, in *1990 DPF Summer Study on High-energy Physics: Research Directions for the Decade (Snowmass 90)*, pp. 0134–136, 12, 1990.

[219] **CDF** Collaboration, F. Abe et al., *The Topology of three jet events in $\bar{p}p$ collisions at $\sqrt{s} = 1.8$ TeV*, *Phys. Rev. D* **45** (1992) 1448–1458.

[220] **D0** Collaboration, B. Abbott, M. Bhattacharjee, D. Elvira, F. Nang, and H. Weerts, *Fixed cone jet definitions in D0 and R(sep)*, 8, 1997.

[221] M. H. Seymour, *Jet shapes in hadron collisions: Higher orders, resummation and hadronization*, *Nucl. Phys. B* **513** (1998) 269–300, [`hep-ph/9707338`].

[222] S. D. Ellis, R. Kleiss, and W. J. Stirling, *THE STANDARD MODEL AND MISSING E(T) OR THE MANY ROADS TO PARADISE*, in *5th Topical Workshop on Proton Antiproton Collider Physics*, 5, 1985.

[223] G. Lejeune Dirichlet, *Über die reduction der positiven quadratischen formen mit drei unbestimmten ganzen zahlen.*, *Journal für die reine und angewandte Mathematik (Crelles Journal)* **1850** (1850), no. 40 209–227.

[224] G. Voronoi, *Nouvelles applications des paramètres continus à la théorie des formes quadratiques. premier mémoire. sur quelques propriétés des formes quadratiques positives parfaites.*, *Journal für die reine und angewandte Mathematik (Crelles Journal)* **1908** (1908), no. 133 97–102.

[225] F. Aurenhammer, *Voronoi diagrams–a survey of a fundamental geometric data structure*, *ACM Computing Surveys (CSUR)* **23** (1991), no. 3 345–405.

[226] S. Fortune, *A sweepline algorithm for voronoi diagrams*, in *Proceedings of the second annual symposium on Computational geometry*, pp. 313–322, 1986.

[227] O. Devillers, S. Meiser, and M. Teillaud, *Fully dynamic delaunay triangulation in logarithmic expected time per operation*, *Computational Geometry* **2** (1992), no. 2 55–80.

[228] O. Devillers, *On deletion in delaunay triangulations*, in *Proceedings of the fifteenth annual symposium on Computational geometry*, pp. 181–188, 1999.

[229] M. Cacciari and G. P. Salam, *Dispelling the N^3 myth for the k_t jet-finder*, *Phys. Lett. B* **641** (2006) 57–61, [`hep-ph/0512210`].

[230] G. C. Blazey et al., *Run II jet physics*, in *Physics at Run II: QCD and Weak Boson Physics Workshop: Final General Meeting*, pp. 47–77, 5, 2000. `hep-ex/0005012`.

[231] M. Dasgupta and G. P. Salam, *Resummation of nonglobal QCD observables*, *Phys. Lett. B* **512** (2001) 323–330, [`hep-ph/0104277`].

[232] A. Banfi, G. Marchesini, and G. Smye, *Away from jet energy flow*, *JHEP* **08** (2002) 006, [`hep-ph/0206076`].

[233] H. Weigert, *Nonglobal jet evolution at finite N(c)*, *Nucl. Phys. B* **685** (2004) 321–350, [`hep-ph/0312050`].

[234] Y. Hatta and T. Ueda, *Resummation of non-global logarithms at finite N_c*, *Nucl. Phys. B* **874** (2013) 808–820, [`arXiv:1304.6930`].

[235] S. Caron-Huot, *Resummation of non-global logarithms and the BFKL equation*, *JHEP* **03** (2018) 036, [`arXiv:1501.03754`].

[236] A. J. Larkoski, I. Moult, and D. Neill, *Non-Global Logarithms, Factorization, and the Soft Substructure of Jets*, *JHEP* **09** (2015) 143, [`arXiv:1501.04596`].

[237] T. Becher, M. Neubert, L. Rothen, and D. Y. Shao, *Effective Field Theory for Jet Processes*, *Phys. Rev. Lett.* **116** (2016), no. 19 192001, [`arXiv:1508.06645`].

[238] R. Ángeles Martínez, M. De Angelis, J. R. Forshaw, S. Plätzer, and M. H. Seymour, *Soft gluon evolution and non-global logarithms*, *JHEP* **05** (2018) 044, [`arXiv:1802.08531`].

[239] A. J. Larkoski, D. Neill, and J. Thaler, *Jet Shapes with the Broadening Axis*, *JHEP* **04** (2014) 017, [`arXiv:1401.2158`].

[240] G. Monge, *Mémoire sur la théorie des déblais et des remblais*, *Mem. Math. Phys. Acad. Royale Sci.* (1781) 666–704.

[241] C. Dupin, "Applications de la géométrie et de la mécanique. 1822. re-edition by bachelier."

[242] P. Appell, *Mémoire sur les déblais et les remblais des systemes continus ou discontinus*, *Mémoires présentes par divers Savants à l'Académie des Sciences de l'Institut de France* **29** (1887) 1–208.

[243] L. V. Kantorovich, *The mathematical method of production planning and organization*, *Management Science* **6** (1939), no. 4 363–422.

[244] L. N. Vaserstein, *Markov processes over denumerable products of spaces, describing large systems of automata*, *Problemy Peredachi Informatsii* **5** (1969), no. 3 64–72.

[245] R. L. Dobrushin, *Prescribing a system of random variables by conditional distributions*, *Theory of Probability & Its Applications* **15** (1970), no. 3 458–486.

[246] F. Ollivier, *Looking for the order of a system of arbitrary ordinary differential equations: De investigando ordine systematis æquationibus differentialium vulgarium cujuscunque*, *Applicable Algebra in Engineering, Communication and Computing* **20** (2009) 7–32.

[247] F. Ollivier, *The reduction to normal form of a non-normal system of differential equations: De æquationum differentialium systemate non normali ad formam normalem revocando*, *Applicable Algebra in Engineering, Communication and Computing* **20** (2009) 33–64.

[248] H. W. Kuhn, *The hungarian method for the assignment problem*, *Naval research logistics quarterly* **2** (1955), no. 1-2 83–97.

[249] P. T. Komiske, E. M. Metodiev, and J. Thaler, *Metric Space of Collider Events*, *Phys. Rev. Lett.* **123** (2019), no. 4 041801, [`arXiv:1902.02346`].

[250] P. T. Komiske, E. M. Metodiev, and J. Thaler, *The Hidden Geometry of Particle Collisions*, *JHEP* **07** (2020) 006, [`arXiv:2004.04159`].

[251] T. Cai, J. Cheng, N. Craig, and K. Craig, *Linearized optimal transport for collider events*, *Phys. Rev. D* **102** (2020), no. 11 116019, [`arXiv:2008.08604`].

[252] T. Cai, J. Cheng, K. Craig, and N. Craig, *Which metric on the space of collider events?*, *Phys. Rev. D* **105** (2022), no. 7 076003, [`arXiv:2111.03670`].

[253] A. J. Larkoski and J. Thaler, *A spectral metric for collider geometry*, *JHEP* **08** (2023) 107, [`arXiv:2305.03751`].

[254] C. Villani et al., *Optimal transport: old and new*, vol. 338. Springer, 2009.

[255] F. Otto and C. Villani, *Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality*, *Journal of Functional Analysis* **173** (2000), no. 2 361–400.

[256] J. Lott and C. Villani, *Ricci curvature for metric-measure spaces via optimal transport*, *Annals of Mathematics* (2009) 903–991.

[257] J. M. Butterworth, J. P. Couchman, B. E. Cox, and B. M. Waugh, *KtJet: A C++ implementation of the K-perpendicular clustering algorithm*, *Comput. Phys. Commun.* **153** (2003) 85–96, [`hep-ph/0210022`].

[258] S. Catani and M. H. Seymour, *A General algorithm for calculating jet cross-sections in NLO QCD*, *Nucl. Phys. B* **485** (1997) 291–419, [`hep-ph/9605323`]. [Erratum: Nucl.Phys.B 510, 503–504 (1998)].

[259] D. Bertolini, T. Chan, and J. Thaler, *Jet Observables Without Jet Algorithms*, *JHEP* **04** (2014) 013, [`arXiv:1310.7584`].

[260] G. Salam, E^∞ *Scheme, unpublished.*

[261] A. J. Larkoski and J. Thaler, *Aspects of jets at 100 TeV*, *Phys. Rev. D* **90** (2014), no. 3 034010, [`arXiv:1406.7011`].

[262] A. Gehrmann-De Ridder and E. W. N. Glover, *A Complete O (alpha alpha-s) calculation of the photon + 1 jet rate in e+ e- annihilation*, *Nucl. Phys. B* **517** (1998) 269–323, [`hep-ph/9707224`].

[263] M. Ritzmann and W. J. Waalewijn, *Fragmentation in Jets at NNLO*, *Phys. Rev. D* **90** (2014), no. 5 054029, [`arXiv:1407.3272`].

[264] P. Cal, D. Neill, F. Ringer, and W. J. Waalewijn, *Calculating the angle between jet axes*, *JHEP* **04** (2020) 211, [`arXiv:1911.06840`].

[265] T.-J. Hou et al., *New CTEQ global analysis of quantum chromodynamics with high-precision data from the LHC*, *Phys. Rev. D* **103** (2021), no. 1 014013, [`arXiv:1912.10053`].

[266] **NNPDF** Collaboration, R. D. Ball et al., *The path to proton structure at 1% accuracy*, *Eur. Phys. J. C* **82** (2022), no. 5 428, [`arXiv:2109.02653`].

[267] D. de Florian, R. Sassot, and M. Stratmann, *Global analysis of fragmentation functions for pions and kaons and their uncertainties*, *Phys. Rev. D* **75** (2007) 114010, [`hep-ph/0703242`].

[268] D. de Florian, R. Sassot, and M. Stratmann, *Global analysis of fragmentation functions for protons and charged hadrons*, *Phys. Rev. D* **76** (2007) 074033, [`arXiv:0707.1506`].

[269] S. Albino, B. A. Kniehl, and G. Kramer, *AKK Update: Improvements from New Theoretical Input and Experimental Data*, *Nucl. Phys. B* **803** (2008) 42–104, [`arXiv:0803.2768`].

[270] D. de Florian, R. Sassot, M. Epele, R. J. Hernández-Pinto, and M. Stratmann, *Parton-to-Pion Fragmentation Reloaded*, *Phys. Rev. D* **91** (2015), no. 1 014035, [`arXiv:1410.6027`].

[271] D. de Florian, M. Epele, R. J. Hernandez-Pinto, R. Sassot, and M. Stratmann, *Parton-to-Kaon Fragmentation Revisited*, *Phys. Rev. D* **95** (2017), no. 9 094019, [`arXiv:1702.06353`].

[272] I. Borsa, D. de Florian, R. Sassot, and M. Stratmann, *Pion fragmentation functions at high energy colliders*, *Phys. Rev. D* **105** (2022), no. 3 L031502, [`arXiv:2110.14015`].

[273] A. Buckley, H. Hoeth, H. Lacker, H. Schulz, and J. E. von Seggern, *Systematic event generator tuning for the LHC*, *Eur. Phys. J. C* **65** (2010) 331–357, [`arXiv:0907.2973`].

[274] J. Bellm, S. Plätzer, P. Richardson, A. Siódmok, and S. Webster, *Reweighting Parton Showers*, *Phys. Rev. D* **94** (2016), no. 3 034028, [`arXiv:1605.08256`].

[275] S. Mrenna and P. Skands, *Automated Parton-Shower Variations in Pythia 8*, *Phys. Rev. D* **94** (2016), no. 7 074005, [arXiv:1605.08352].

[276] E. Bothmann, M. Schönherr, and S. Schumann, *Reweighting QCD matrix-element and parton-shower calculations*, *Eur. Phys. J. C* **76** (2016), no. 11 590, [arXiv:1606.08753].

[277] J. Bellm and L. Gellersen, *High dimensional parameter tuning for event generators*, *Eur. Phys. J. C* **80** (2020), no. 1 54, [arXiv:1908.10811].

[278] M. Krishnamoorthy, H. Schulz, X. Ju, W. Wang, S. Leyffer, Z. Marshall, S. Mrenna, J. Müller, and J. B. Kowalkowski, *Apprentice for Event Generator Tuning*, *EPJ Web Conf.* **251** (2021) 03060, [arXiv:2103.05748].

[279] C. Bierlich, P. Ilten, T. Menzo, S. Mrenna, M. Szewc, M. K. Wilkinson, A. Youssef, and J. Zupan, *Reweighting Monte Carlo Predictions and Automated Fragmentation Variations in Pythia 8*, arXiv:2308.13459.

[280] **CMS** Collaboration, A. M. Sirunyan et al., *Search for physics beyond the standard model in events with jets and two same-sign or at least three charged leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV*, *Eur. Phys. J. C* **80** (2020), no. 8 752, [arXiv:2001.10086].

[281] **CMS** Collaboration, A. M. Sirunyan et al., *Search for supersymmetry in proton-proton collisions at $\sqrt{s} = 13$ TeV in events with high-momentum Z bosons and missing transverse momentum*, *JHEP* **09** (2020) 149, [arXiv:2008.04422].

[282] **CMS** Collaboration, A. M. Sirunyan et al., *Search for supersymmetry in final states with two oppositely charged same-flavor leptons and missing transverse momentum in proton-proton collisions at $\sqrt{s} = 13$ TeV*, *JHEP* **04** (2021) 123, [arXiv:2012.08600].

[283] **ATLAS** Collaboration, G. Aad et al., *Search for supersymmetry in events with four or more charged leptons in 139 fb$^{?1}$ of $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector*, *JHEP* **07** (2021) 167, [arXiv:2103.11684].

[284] **ATLAS** Collaboration, G. Aad et al., *Search for charginos and neutralinos in final states with two boosted hadronically decaying bosons and missing transverse momentum in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Phys. Rev. D* **104** (2021), no. 11 112010, [arXiv:2108.07586].

[285] **CMS** Collaboration, A. Tumasyan et al., *Search for higgsinos decaying to two Higgs bosons and missing transverse momentum in proton-proton collisions at $\sqrt{s} = 13$ TeV*, *JHEP* **05** (2022) 014, [arXiv:2201.04206].

[286] **CMS** Collaboration, A. Tumasyan et al., *Search for light Higgs bosons from supersymmetric cascade decays in pp collisions at $\sqrt{s} = 13$ TeV*, *Eur. Phys. J. C* **83** (2023), no. 7 571, [arXiv:2204.13532].

[287] **ATLAS** Collaboration, G. Aad et al., *Search for pair production of higgsinos in events with two Higgs bosons and missing transverse momentum in $\sqrt{s} = 13$ TeV pp collisions at the ATLAS experiment*, arXiv:2401.14922.

[288] L. M. Jones, *Tests for Determining the Parton Ancestor of a Hadron Jet*, *Phys. Rev. D* **39** (1989) 2550.

[289] Z. Fodor, *How to See the Differences Between Quark and Gluon Jets*, *Phys. Rev. D* **41** (1990) 1726.

[290] L. Lonnblad, C. Peterson, and T. Rognvaldsson, *Finding Gluon Jets With a Neural Trigger*, *Phys. Rev. Lett.* **65** (1990) 1321–1324.

[291] L. Lonnblad, C. Peterson, and T. Rognvaldsson, *Using neural networks to identify jets*, *Nucl. Phys. B* **349** (1991) 675–702.

[292] I. Csabai, F. Czako, and Z. Fodor, *Quark and gluon jet separation using neural networks*, *Phys. Rev. D* **44** (1991) 1905–1908.

[293] L. Jones, *TOWARDS A SYSTEMATIC JET CLASSIFICATION*, *Phys. Rev. D* **42** (1990) 811–814.

[294] J. Pumplin, *How to tell quark jets from gluon jets*, *Phys. Rev. D* **44** (1991) 2025–2032.

[295] **OPAL** Collaboration, P. D. Acton et al., *A Study of differences between quark and gluon jets using vertex tagging of quark jets*, *Z. Phys. C* **58** (1993) 387–404.

[296] J. Gallicchio and M. D. Schwartz, *Quark and Gluon Tagging at the LHC*, *Phys. Rev. Lett.* **107** (2011) 172001, [`arXiv:1106.3076`].

[297] J. Gallicchio and M. D. Schwartz, *Quark and Gluon Jet Substructure*, *JHEP* **04** (2013) 090, [`arXiv:1211.7038`].

[298] A. J. Larkoski, G. P. Salam, and J. Thaler, *Energy Correlation Functions for Jet Substructure*, *JHEP* **06** (2013) 108, [`arXiv:1305.0007`].

[299] P. Gras, S. Höche, D. Kar, A. Larkoski, L. Lönnblad, S. Plätzer, A. Siódmok, P. Skands, G. Soyez, and J. Thaler, *Systematics of quark/gluon tagging*, *JHEP* **07** (2017) 091, [`arXiv:1704.03878`].

[300] A. Banfi, G. P. Salam, and G. Zanderighi, *Infrared safe definition of jet flavor*, *Eur. Phys. J. C* **47** (2006) 113–124, [`hep-ph/0601139`].

[301] M. J. Strassler, *An Unorthodox introduction to supersymmetric gauge theory*, in *Theoretical Advanced Study Institute in Elementary Particle Physics (TASI 2001): Strings, Branes and EXTRA Dimensions*, pp. 561–638, 9, 2003. `hep-th/0309149`.

[302] E. P. Wigner, *On the quantum correction for thermodynamic equilibrium*, *Phys. Rev.* **40** (1932) 749–760.

[303] P. A. M. Dirac, *Bakerian lecture-the physical interpretation of quantum mechanics*, *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **180** (1942), no. 980 1–40.

[304] R. P. Feynman, *NEGATIVE PROBABILITY*, . PRINT-84-0846 (CAL-TECH).

[305] B. Andersson, G. Gustafson, L. Lonnblad, and U. Pettersson, *Coherence Effects in Deep Inelastic Scattering*, *Z. Phys. C* **43** (1989) 625.

[306] J. D. Bjorken, *A Plumber's view of perturbative QCD*, *Phys. Rev. D* **45** (1992) 4077–4087.

[307] F. A. Dreyer, G. P. Salam, and G. Soyez, *The Lund Jet Plane*, *JHEP* **12** (2018) 064, [`arXiv:1807.04758`].

[308] F. A. Dreyer and H. Qu, *Jet tagging in the Lund plane with graph networks*, *JHEP* **03** (2021) 052, [`arXiv:2012.08526`].

[309] **ATLAS** Collaboration, G. Aad et al., *Measurement of the Lund Jet Plane Using Charged Particles in 13 TeV Proton-Proton Collisions with the ATLAS Detector*, *Phys. Rev. Lett.* **124** (2020), no. 22 222002, [`arXiv:2004.03540`].

[310] A. Lifson, G. P. Salam, and G. Soyez, *Calculating the primary Lund Jet Plane density*, *JHEP* **10** (2020) 170, [`arXiv:2007.06578`].

[311] **ALICE** Collaboration, *Physics Preliminary Summary: Measurement of the primary Lund plane density in pp collisions at $\sqrt{s} = 13$ TeV with ALICE*, . ALICE-PUBLIC-2021-002.

[312] F. A. Dreyer, G. Soyez, and A. Takacs, *Quarks and gluons in the Lund plane*, *JHEP* **08** (2022) 177, [`arXiv:2112.09140`].

[313] **CMS** Collaboration, A. Hayrapetyan et al., *Measurement of the primary Lund jet plane density in proton-proton collisions at $\sqrt{s} = 13$ TeV*, `arXiv:2312.16343`.

[314] C. Bierlich et al., *A comprehensive guide to the physics and usage of PYTHIA 8.3*, *SciPost Phys. Codeb.* **2022** (2022) 8, [`arXiv:2203.11601`].

[315] M. Bahr et al., *Herwig++ Physics and Manual*, *Eur. Phys. J. C* **58** (2008) 639–707, [`arXiv:0803.0883`].

[316] J. Bellm et al., *Herwig 7.0/Herwig++ 3.0 release note*, *Eur. Phys. J. C* **76** (2016), no. 4 196, [`arXiv:1512.01178`].

[317] **Sherpa** Collaboration, E. Bothmann et al., *Event Generation with Sherpa 2.2*, *SciPost Phys.* **7** (2019), no. 3 034, [`arXiv:1905.09127`].

[318] K. Hamilton, R. Medves, G. P. Salam, L. Scyboz, and G. Soyez, *Colour and logarithmic accuracy in final-state parton showers*, *JHEP* **03** (2021), no. 041 041, [`arXiv:2011.10054`].

[319] J. R. Forshaw, J. Holguin, and S. Plätzer, *Building a consistent parton shower*, *JHEP* **09** (2020) 014, [`arXiv:2003.06400`].

[320] Z. Nagy and D. E. Soper, *Summations of large logarithms by parton showers*, *Phys. Rev. D* **104** (2021), no. 5 054049, [`arXiv:2011.04773`].

[321] F. Herren, S. Höche, F. Krauss, D. Reichelt, and M. Schoenherr, *A new approach to color-coherent parton evolution*, `arXiv:2208.06057`.

[322] C. T. Preuss, *A partitioned dipole-antenna shower with improved transverse recoil*, `arXiv:2403.19452`.

[323] C. F. Berger, T. Kucs, and G. F. Sterman, *Event shape / energy flow correlations*, *Phys. Rev. D* **68** (2003) 014012, [`hep-ph/0303051`].

[324] L. G. Almeida, S. J. Lee, G. Perez, G. F. Sterman, I. Sung, and J. Virzi, *Substructure of high-p_T Jets at the LHC*, *Phys. Rev. D* **79** (2009) 074017, [`arXiv:0807.0234`].

[325] S. D. Ellis, C. K. Vermilion, J. R. Walsh, A. Hornig, and C. Lee, *Jet Shapes and Jet Algorithms in SCET*, *JHEP* **11** (2010) 101, [`arXiv:1001.0014`].

[326] S. Brandt, C. Peyrou, R. Sosnowski, and A. Wroblewski, *The Principal axis of jets. An Attempt to analyze high-energy collisions as two-body processes*, *Phys. Lett.* **12** (1964) 57–61.

[327] E. Farhi, *A QCD Test for Jets*, *Phys. Rev. Lett.* **39** (1977) 1587–1588.

[328] L. Clavelli, *Jet Invariant Mass in Quantum Chromodynamics*, *Phys. Lett. B* **85** (1979) 111–114.

[329] P. E. L. Rakow and B. R. Webber, *Transverse Momentum Moments of Hadron Distributions in QCD Jets*, *Nucl. Phys. B* **191** (1981) 63–74.

[330] R. K. Ellis and B. R. Webber, *QCD Jet Broadening in Hadron Hadron Collisions*, *Conf. Proc. C* **860623** (1986) 74.

[331] S. Catani, G. Turnock, and B. R. Webber, *Jet broadening measures in e^+e^- annihilation*, *Phys. Lett. B* **295** (1992) 269–276.

[332] Y. L. Dokshitzer, A. Lucenti, G. Marchesini, and G. P. Salam, *On the QCD analysis of jet broadening*, *JHEP* **01** (1998) 011, [`hep-ph/9801324`].

[333] H. Georgi and M. Machacek, *A Simple QCD Prediction of Jet Structure in e+ e- Annihilation*, *Phys. Rev. Lett.* **39** (1977) 1237.

[334] G. Parisi, *Super Inclusive Cross-Sections*, *Phys. Lett. B* **74** (1978) 65–67.

[335] J. F. Donoghue, F. E. Low, and S.-Y. Pi, *Tensor Analysis of Hadronic Jets in Quantum Chromodynamics*, *Phys. Rev. D* **20** (1979) 2759.

[336] A. Banfi, G. P. Salam, and G. Zanderighi, *Principles of general final-state resummation and automated implementation*, *JHEP* **03** (2005) 073, [`hep-ph/0407286`].

[337] I. Moult, L. Necib, and J. Thaler, *New Angles on Energy Correlation Functions*, *JHEP* **12** (2016) 153, [`arXiv:1609.07483`].

[338] C. L. Basham, L. S. Brown, S. D. Ellis, and S. T. Love, *Energy Correlations in electron - Positron Annihilation: Testing QCD*, *Phys. Rev. Lett.* **41** (1978) 1585.

[339] D. M. Hofman and J. Maldacena, *Conformal collider physics: Energy and charge correlations*, *JHEP* **05** (2008) 012, [`arXiv:0803.1467`].

[340] L. J. Dixon, M.-X. Luo, V. Shtabovenko, T.-Z. Yang, and H. X. Zhu, *Analytical Computation of Energy-Energy Correlation at Next-to-Leading Order in QCD*, *Phys. Rev. Lett.* **120** (2018), no. 10 102001, [`arXiv:1801.03219`].

[341] L. J. Dixon, I. Moult, and H. X. Zhu, *Collinear limit of the energy-energy correlator*, *Phys. Rev. D* **100** (2019), no. 1 014009, [`arXiv:1905.01310`].

[342] J. M. Henn, E. Sokatchev, K. Yan, and A. Zhiboedov, *Energy-energy correlation in N=4 super Yang-Mills theory at next-to-next-to-leading order*, *Phys. Rev. D* **100** (2019), no. 3 036010, [`arXiv:1903.05314`].

[343] H. Chen, I. Moult, and H. X. Zhu, *Quantum Interference in Jet Substructure from Spinning Gluons*, *Phys. Rev. Lett.* **126** (2021), no. 11 112003, [`arXiv:2011.02492`].

[344] P. T. Komiske, I. Moult, J. Thaler, and H. X. Zhu, *Analyzing N-Point Energy Correlators inside Jets with CMS Open Data*, *Phys. Rev. Lett.* **130** (2023), no. 5 051901, [`arXiv:2201.07800`].

[345] **CMS** Collaboration, A. Hayrapetyan et al., *Measurement of energy correlators inside jets and determination of the strong coupling $\alpha_{\mathrm{S}}(m_{\mathrm{Z}})$*, `arXiv:2402.13864`.

[346] V. V. Sudakov, *Vertex parts at very high-energies in quantum electrodynamics*, *Sov. Phys. JETP* **3** (1956) 65–71.

[347] S. Catani, L. Trentadue, G. Turnock, and B. R. Webber, *Resummation of large logarithms in e+ e- event shape distributions*, *Nucl. Phys. B* **407** (1993) 3–42.

[348] A. M. Polyakov, *Gauge Fields as Rings of Glue*, *Nucl. Phys. B* **164** (1980) 171–188.

[349] G. P. Korchemsky and A. V. Radyushkin, *Renormalization of the Wilson Loops Beyond the Leading Order*, *Nucl. Phys. B* **283** (1987) 342–364.

[350] D. Correa, J. Henn, J. Maldacena, and A. Sever, *The cusp anomalous dimension at three loops and beyond*, *JHEP* **05** (2012) 098, [arXiv:1203.1019].

[351] A. Grozin, J. M. Henn, G. P. Korchemsky, and P. Marquard, *Three Loop Cusp Anomalous Dimension in QCD*, *Phys. Rev. Lett.* **114** (2015), no. 6 062006, [arXiv:1409.0023].

[352] A. Grozin, J. M. Henn, G. P. Korchemsky, and P. Marquard, *The three-loop cusp anomalous dimension in QCD and its supersymmetric extensions*, *JHEP* **01** (2016) 140, [arXiv:1510.07803].

[353] S. Moch, B. Ruijl, T. Ueda, J. A. M. Vermaseren, and A. Vogt, *On quartic colour factors in splitting functions and the gluon cusp anomalous dimension*, *Phys. Lett. B* **782** (2018) 627–632, [arXiv:1805.09638].

[354] J. M. Henn, G. P. Korchemsky, and B. Mistlberger, *The full four-loop cusp anomalous dimension in $\mathcal{N} = 4$ super Yang-Mills and QCD*, *JHEP* **04** (2020) 018, [arXiv:1911.10174].

[355] A. von Manteuffel, E. Panzer, and R. M. Schabinger, *Cusp and collinear anomalous dimensions in four-loop QCD from form factors*, *Phys. Rev. Lett.* **124** (2020), no. 16 162001, [arXiv:2002.04617].

[356] **ATLAS** Collaboration, G. 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** (2012) 1–29, [arXiv:1207.7214].

[357] **CMS** Collaboration, S. Chatrchyan et al., *Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC*, *Phys. Lett. B* **716** (2012) 30–61, [arXiv:1207.7235].

[358] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N-subjettiness*, *JHEP* **03** (2011) 015, [arXiv:1011.2268].

[359] J. Thaler and K. Van Tilburg, *Maximizing Boosted Top Identification by Minimizing N-subjettiness*, *JHEP* **02** (2012) 093, [arXiv:1108.2701].

[360] S. Brandt and H. D. Dahmen, *Axes and Scalar Measures of Two-Jet and Three-Jet Events*, *Z. Phys. C* **1** (1979) 61.

[361] I. W. Stewart, F. J. Tackmann, and W. J. Waalewijn, *N-Jettiness: An Inclusive Event Shape to Veto Jets*, *Phys. Rev. Lett.* **105** (2010) 092002, [arXiv:1004.2489].

[362] J.-H. Kim, *Rest Frame Subjet Algorithm With SISCone Jet For Fully Hadronic Decaying Higgs Search*, *Phys. Rev. D* **83** (2011) 011502, [arXiv:1011.1493].

[363] G. Kasieczka, S. Marzani, G. Soyez, and G. Stagnitto, *Towards Machine Learning Analytics for Jet Substructure*, *JHEP* **09** (2020) 195, [arXiv:2007.04319].

[364] A. J. Larkoski and J. Thaler, *Unsafe but Calculable: Ratios of Angularities in Perturbative QCD*, *JHEP* **09** (2013) 137, [arXiv:1307.1699].

[365] **ALEPH** Collaboration, A. Heister et al., *Studies of QCD at e+ e- centre-of-mass energies between 91-GeV and 209-GeV*, *Eur. Phys. J. C* **35** (2004) 457–486.

[366] D. E. Kaplan, K. Rehermann, M. D. Schwartz, and B. Tweedie, *Top Tagging: A Method for Identifying Boosted Hadronically Decaying Top Quarks*, *Phys. Rev. Lett.* **101** (2008) 142001, [arXiv:0806.0848].

[367] **ATLAS** Collaboration, A. Airapetian et al., *ATLAS: Detector and physics performance technical design report. Volume 2*, 5, 1999.

[368] **CMS** Collaboration, G. L. Bayatian et al., *CMS technical design report, volume II: Physics performance*, *J. Phys. G* **34** (2007), no. 6 995–1579.

[369] M. Cacciari, G. P. Salam, and G. Soyez, *The Catchment Area of Jets*, *JHEP* **04** (2008) 005, [arXiv:0802.1188].

[370] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, *Jet substructure as a new Higgs search channel at the LHC*, *Phys. Rev. Lett.* **100** (2008) 242001, [arXiv:0802.2470].

[371] M. H. Seymour, *Searches for new particles using cone and cluster jet algorithms: A Comparative study*, *Z. Phys. C* **62** (1994) 127–138.

[372] J. M. Butterworth, B. E. Cox, and J. R. Forshaw, *WW scattering at the CERN LHC*, *Phys. Rev. D* **65** (2002) 096014, [hep-ph/0201098].

[373] J. M. Butterworth, J. R. Ellis, and A. R. Raklev, *Reconstructing sparticle mass spectra using hadronic decays*, *JHEP* **05** (2007) 033, [hep-ph/0702150].

[374] **ATLAS** Collaboration, M. Aaboud et al., *Observation of $H \to b\bar{b}$ decays and VH production with the ATLAS detector*, *Phys. Lett. B* **786** (2018) 59–86, [arXiv:1808.08238].

[375] **CMS** Collaboration, A. M. Sirunyan et al., *Observation of Higgs boson decay to bottom quarks*, *Phys. Rev. Lett.* **121** (2018), no. 12 121801, [arXiv:1808.08242].

[376] J. Lin, M. Freytsis, I. Moult, and B. Nachman, *Boosting $H \to b\bar{b}$ with Machine Learning*, *JHEP* **10** (2018) 101, [arXiv:1807.10768].

[377] K. Datta, A. Larkoski, and B. Nachman, *Automating the Construction of Jet Observables with Machine Learning*, *Phys. Rev. D* **100** (2019), no. 9 095016, [arXiv:1902.07180].

[378] E. A. Moreno, T. Q. Nguyen, J.-R. Vlimant, O. Cerri, H. B. Newman, A. Periwal, M. Spiropulu, J. M. Duarte, and M. Pierini, *Interaction networks for the identification of boosted $H \to b\bar{b}$ decays*, *Phys. Rev. D* **102** (2020), no. 1 012010, [arXiv:1909.12285].

[379] A. Chakraborty, S. H. Lim, and M. M. Nojiri, *Interpretable deep learning for two-prong jet classification with jet spectra*, *JHEP* **07** (2019) 135, [arXiv:1904.02092].

[380] **CMS** Collaboration, A. M. Sirunyan et al., *Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques*, *JINST* **15** (2020), no. 06 P06005, [arXiv:2004.08262].

[381] Y.-L. Chung, S.-C. Hsu, and B. Nachman, *Disentangling Boosted Higgs Boson Production Modes with Machine Learning*, *JINST* **16** (2021) P07002, [arXiv:2009.05930].

[382] B. Tannenwald, C. Neu, A. Li, G. Buehlmann, A. Cuddeback, L. Hatfield, R. Parvatam, and C. Thompson, *Benchmarking Machine Learning Techniques with Di-Higgs Production at the LHC*, `arXiv:2009.06754`.

[383] J. Guo, J. Li, T. Li, and R. Zhang, *Boosted Higgs boson jet reconstruction via a graph neural network*, *Phys. Rev. D* **103** (2021), no. 11 116025, [`arXiv:2010.05464`].

[384] M. Abbas, A. Khan, A. S. Qureshi, and M. W. Khan, *Extracting Signals of Higgs Boson From Background Noise Using Deep Neural Networks*, `arXiv:2010.08201`.

[385] C. Jang, S.-K. Ko, J. Choi, J. Lim, Y.-K. Noh, and T. J. Kim, *Learning to increase matching efficiency in identifying additional b-jets in the $t\bar{t}b\bar{b}$ process*, *Eur. Phys. J. Plus* **137** (2022), no. 7 870, [`arXiv:2103.09129`].

[386] C. K. Khosa and S. Marzani, *Higgs boson tagging with the Lund jet plane*, *Phys. Rev. D* **104** (2021), no. 5 055043, [`arXiv:2105.03989`].

[387] **ATLAS** Collaboration, G. Aad et al., *ATLAS flavour-tagging algorithms for the LHC Run 2 pp collision dataset*, *Eur. Phys. J. C* **83** (2023), no. 7 681, [`arXiv:2211.16345`].

[388] **CMS** Collaboration, *Performance of heavy-flavour jet identification in boosted topologies in proton-proton collisions at $\sqrt{s} = 13$ TeV*, 2023.
CMS-PAS-BTV-22-001.

[389] A. Butter et al., *The Machine Learning landscape of top taggers*, *SciPost Phys.* **7** (2019) 014, [`arXiv:1902.09914`].

[390] A. J. Larkoski, *Binary discrimination through next-to-leading order*, *JHEP* **03** (2024) 057, [`arXiv:2309.14417`].

[391] A. J. Larkoski, *General analysis for observing quantum interference at colliders*, *Phys. Rev. D* **105** (2022), no. 9 096012, [`arXiv:2201.03159`].

[392] K. Gottfried and J. D. Jackson, *On the Connection between production mechanism and decay of resonances at high-energies*, *Nuovo Cim.* **33** (1964) 309–330.

[393] B. R. Webber, *Monte Carlo Simulation of Hard Hadronic Processes*, *Ann. Rev. Nucl. Part. Sci.* **36** (1986) 253–286.

[394] J. C. Collins, *Spin Correlations in Monte Carlo Event Generators*, *Nucl. Phys. B* **304** (1988) 794–804.

[395] I. G. Knowles, *Angular Correlations in QCD*, *Nucl. Phys. B* **304** (1988) 767–793.

[396] M. R. Buckley, H. Murayama, W. Klemm, and V. Rentala, *Discriminating spin through quantum interference*, *Phys. Rev. D* **78** (2008) 014028, [`arXiv:0711.0364`].

[397] M. R. Buckley, B. Heinemann, W. Klemm, and H. Murayama, *Quantum Interference Effects Among Helicities at LEP-II and Tevatron*, *Phys. Rev. D* **77** (2008) 113017, [`arXiv:0804.0476`].

[398] A. Alves, O. J. P. Eboli, M. C. Gonzalez-Garcia, and J. K. Mizukoshi, *Deciphering the spin of new resonances in Higgsless models*, *Phys. Rev. D* **79** (2009) 035009, [`arXiv:0810.1952`].

[399] F. Boudjema and R. K. Singh, *A Model independent spin analysis of fundamental particles using azimuthal asymmetries*, *JHEP* **07** (2009) 028, [`arXiv:0903.4705`].

[400] H. Murayama and V. Rentala, *Randall-Sundrum graviton spin determination using azimuthal angular dependence*, *Phys. Rev. D* **85** (2012) 095005, [`arXiv:0904.4561`].

[401] A. Karlberg, G. P. Salam, L. Scyboz, and R. Verheyen, *Spin correlations in final-state parton showers and jet observables*, *Eur. Phys. J. C* **81** (2021), no. 8 681, [`arXiv:2103.16526`].

[402] K. Hamilton, A. Karlberg, G. P. Salam, L. Scyboz, and R. Verheyen, *Soft spin correlations in final-state parton showers*, *JHEP* **03** (2022) 193, [`arXiv:2111.01161`].

[403] A. J. Larkoski, *Designing observables for quantum interference in jets at subleading color*, *SciPost Phys.* **14** (2023), no. 3 041, [`arXiv:2205.12375`].

[404] S. Caletti, A. J. Larkoski, S. Marzani, and D. Reichelt, *Practical jet flavour through NNLO*, *Eur. Phys. J. C* **82** (2022), no. 7 632, [`arXiv:2205.01109`].

[405] S. Caletti, A. J. Larkoski, S. Marzani, and D. Reichelt, *A fragmentation approach to jet flavor*, *JHEP* **10** (2022) 158, [`arXiv:2205.01117`].

[406] M. Czakon, A. Mitov, and R. Poncelet, *Infrared-safe flavoured anti-k_T jets*, *JHEP* **04** (2023) 138, [`arXiv:2205.11879`].

[407] R. Gauld, A. Huss, and G. Stagnitto, *Flavor Identification of Reconstructed Hadronic Jets*, *Phys. Rev. Lett.* **130** (2023), no. 16 161901, [`arXiv:2208.11138`].

[408] F. Caola, R. Grabarczyk, M. L. Hutt, G. P. Salam, L. Scyboz, and J. Thaler, *Flavored jets with exact anti-kt kinematics and tests of infrared and collinear safety*, *Phys. Rev. D* **108** (2023), no. 9 094010, [`arXiv:2306.07314`].

[409] A. Buckley, G. Callea, A. J. Larkoski, and S. Marzani, *An Optimal Observable for Color Singlet Identification*, *SciPost Phys.* **9** (2020) 026, [`arXiv:2006.10480`].

[410] J. Thaler and L.-T. Wang, *Strategies to Identify Boosted Tops*, *JHEP* **07** (2008) 092, [`arXiv:0806.0023`].

[411] J. Gallicchio and M. D. Schwartz, *Seeing in Color: Jet Superstructure*, *Phys. Rev. Lett.* **105** (2010) 022001, [`arXiv:1001.5027`].

[412] A. Hook, M. Jankowiak, and J. G. Wacker, *Jet Dipolarity: Top Tagging with Color Flow*, *JHEP* **04** (2012) 007, [`arXiv:1102.1012`].

[413] A. J. Larkoski, I. Moult, and D. Neill, *Power Counting to Better Jet Observables*, *JHEP* **12** (2014) 009, [`arXiv:1409.6298`].

[414] J. R. Walsh and S. Zuberi, *Factorization Constraints on Jet Substructure*, `arXiv:1110.5333`.

[415] C. W. Bauer, F. J. Tackmann, J. R. Walsh, and S. Zuberi, *Factorization and Resummation for Dijet Invariant Mass Spectra*, *Phys. Rev. D* **85** (2012) 074006, [`arXiv:1106.6047`].

[416] A. J. Larkoski, I. Moult, and D. Neill, *Toward Multi-Differential Cross Sections: Measuring Two Angularities on a Single Jet*, *JHEP* **09** (2014) 046, [`arXiv:1401.4458`].

[417] M. Procura, W. J. Waalewijn, and L. Zeune, *Resummation of Double-Differential Cross Sections and Fully-Unintegrated Parton Distribution Functions*, *JHEP* **02** (2015) 117, [`arXiv:1410.6483`].

[418] **CMS** Collaboration, J. Olsen, "Cms 13 tev results."
https://indico.cern.ch/event/442432/contributions/1946922/attachments/
1205563/1756687/CMS_13_TeV_results_public.pdf.

[419] **ATLAS** Collaboration, M. Kado, "Results with the full 2015 data sample from the atlas experiment."
https://indico.cern.ch/event/442432/contributions/1946921/attachments/
1205572/1759985/CERN-Seminar.pdf.

[420] **ATLAS** Collaboration, M. Aaboud et al., *Search for resonances in diphoton events at $\sqrt{s}=13$ TeV with the ATLAS detector*, *JHEP* **09** (2016) 001, [arXiv:1606.03833].

[421] **CMS** Collaboration, V. Khachatryan et al., *Search for Resonant Production of High-Mass Photon Pairs in Proton-Proton Collisions at $\sqrt{s} =8$ and 13 TeV*, *Phys. Rev. Lett.* **117** (2016), no. 5 051802, [arXiv:1606.04093].

[422] M. Backović, *A Theory of Ambulance Chasing*, arXiv:1603.01204.

[423] G. Soyez, G. P. Salam, J. Kim, S. Dutta, and M. Cacciari, *Pileup subtraction for jet shapes*, *Phys. Rev. Lett.* **110** (2013), no. 16 162001, [arXiv:1211.2811].

[424] A. J. Larkoski, S. Marzani, and J. Thaler, *Sudakov Safety in Perturbative QCD*, *Phys. Rev. D* **91** (2015), no. 11 111501, [arXiv:1502.01719].

[425] A. J. Larkoski, I. Moult, and D. Neill, *Analytic Boosted Boson Discrimination*, *JHEP* **05** (2016) 117, [arXiv:1507.03018].

[426] M. Dasgupta, L. Schunk, and G. Soyez, *Jet shapes for boosted jet two-prong decays from first-principles*, *JHEP* **04** (2016) 166, [arXiv:1512.00516].

[427] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, *Soft Drop*, *JHEP* **05** (2014) 146, [arXiv:1402.2657].

[428] F. A. Dreyer, L. Necib, G. Soyez, and J. Thaler, *Recursive Soft Drop*, *JHEP* **06** (2018) 093, [arXiv:1804.03657].

[429] P. Cal, K. Lee, F. Ringer, and W. J. Waalewijn, *Jet energy drop*, *JHEP* **11** (2020) 012, [arXiv:2007.12187].

[430] Y. Mehtar-Tani, A. Soto-Ontoso, and K. Tywoniuk, *Dynamical grooming of QCD jets*, *Phys. Rev. D* **101** (2020), no. 3 034004, [arXiv:1911.00375].

[431] S. Alipour-Fard, P. T. Komiske, E. M. Metodiev, and J. Thaler, *Pileup and Infrared Radiation Annihilation (PIRANHA): a paradigm for continuous jet grooming*, *JHEP* **09** (2023) 157, [arXiv:2305.00989].

[432] K. Weierstrass, *Über die analytische darstellbarkeit sogenannter willkürlicher functionen einer reellen veränderlichen*, *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin* **2** (1885) 633–639.

[433] M. H. Stone, *Applications of the theory of boolean rings to general topology*, *Transactions of the American Mathematical Society* **41** (1937), no. 3 375–481.

[434] M. H. Stone, *The generalized weierstrass approximation theorem*, *Mathematics Magazine* **21** (1948), no. 5 237–254.

[435] P. T. Komiske, E. M. Metodiev, and J. Thaler, *Cutting Multiparticle Correlators Down to Size*, *Phys. Rev. D* **101** (2020), no. 3 036019, [arXiv:1911.04491].

[436] P. Cal, J. Thaler, and W. J. Waalewijn, *Power counting energy flow polynomials*, *JHEP* **09** (2022) 021, [`arXiv:2205.06818`].

[437] P. T. Komiske, E. M. Metodiev, and J. Thaler, *Energy Flow Networks: Deep Sets for Particle Jets*, *JHEP* **01** (2019) 121, [`arXiv:1810.05165`].

[438] T. Faucett, J. Thaler, and D. Whiteson, *Mapping Machine-Learned Physics into a Human-Readable Space*, *Phys. Rev. D* **103** (2021), no. 3 036020, [`arXiv:2010.11998`].

[439] A. Romero, D. Whiteson, M. Fenton, J. Collado, and P. Baldi, *Safety of Quark/Gluon Jet Classification*, `arXiv:2103.09103`.

[440] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, *Techniques for improved heavy particle searches with jet substructure*, *Phys. Rev. D* **80** (2009) 051501, [`arXiv:0903.5081`].

[441] S. D. Ellis, C. K. Vermilion, and J. R. Walsh, *Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches*, *Phys. Rev. D* **81** (2010) 094023, [`arXiv:0912.0033`].

[442] L. G. Almeida, S. J. Lee, G. Perez, G. Sterman, and I. Sung, *Template Overlap Method for Massive Jets*, *Phys. Rev. D* **82** (2010) 054034, [`arXiv:1006.2035`].

[443] M. Jankowiak and A. J. Larkoski, *Jet Substructure Without Trees*, *JHEP* **06** (2011) 057, [`arXiv:1104.1646`].

[444] A. J. Larkoski, I. Moult, and D. Neill, *Building a Better Boosted Top Tagger*, *Phys. Rev. D* **91** (2015), no. 3 034035, [`arXiv:1411.0665`].

[445] B. Nachman, "Private communication."

[446] C. Frye, A. J. Larkoski, J. Thaler, and K. Zhou, *Casimir Meets Poisson: Improved Quark/Gluon Discrimination with Counting Observables*, *JHEP* **09** (2017) 083, [`arXiv:1704.06266`].

[447] R. Medves, A. Soto-Ontoso, and G. Soyez, *Lund and Cambridge multiplicities for precision physics*, *JHEP* **10** (2022) 156, [`arXiv:2205.02861`].

[448] R. Medves, A. Soto-Ontoso, and G. Soyez, *Lund multiplicity in QCD jets*, *JHEP* **04** (2023) 104, [`arXiv:2212.05076`].

[449] D. J. Gross and F. Wilczek, *Ultraviolet Behavior of Nonabelian Gauge Theories*, *Phys. Rev. Lett.* **30** (1973) 1343–1346.

[450] H. D. Politzer, *Reliable Perturbative Results for Strong Interactions?*, *Phys. Rev. Lett.* **30** (1973) 1346–1349.

[451] G. 't Hooft, *When was asymptotic freedom discovered? or the rehabilitation of quantum field theory*, *Nucl. Phys. B Proc. Suppl.* **74** (1999) 413–425, [`hep-th/9808154`].

[452] D. Gross and F. Wilczek, *A watershed: the emergence of QCD*, *CERN Courier* **53** (2013), no. 1 24–27.

[453] **Particle Data Group** Collaboration, R. L. Workman and Others, *Review of Particle Physics*, *PTEP* **2022** (2022) 083C01.

[454] D. Amati and G. Veneziano, *Preconfinement as a Property of Perturbative QCD*, *Phys. Lett. B* **83** (1979) 87–92.

[455] A. Bassetto, M. Ciafaloni, and G. Marchesini, *Color Singlet Distributions and Mass Damping in Perturbative QCD*, Phys. Lett. B **83** (1979) 207–212.

[456] R. D. Field and S. Wolfram, *A QCD Model for e+ e- Annihilation*, Nucl. Phys. B **213** (1983) 65–84.

[457] B. R. Webber, *A QCD Model for Jet Fragmentation Including Soft Gluon Interference*, Nucl. Phys. B **238** (1984) 492–528.

[458] A. E. Chudakov *Izv. Akad. Nauk Ser. Fiz.* **19** (1955) 589.

[459] D. Perkins, *Ionization at the origin of electron pairs, and the lifetime of the neutral pion*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **46** (1955), no. 381 1146–1148, [https://doi.org/10.1080/14786441008521131].

[460] F. Hausdorff, *Dimension und äußeres maß*, Mathematische Annalen **79** (1918), no. 1-2 157–179.

[461] B. Mandelbrot, *How long is the coast of britain? statistical self-similarity and fractional dimension*, Science **156** (1967), no. 3775 636–638, [https://www.science.org/doi/pdf/10.1126/science.156.3775.636].

[462] G. Gustafson and A. Nilsson, *Multifractal dimensions in QCD cascades*, Z. Phys. C **52** (1991) 533–542.

[463] A. Bassetto, M. Ciafaloni, and G. Marchesini, *Inelastic Distributions and Color Structure in Perturbative QCD*, Nucl. Phys. B **163** (1980) 477–518.

[464] W. Furmanski, R. Petronzio, and S. Pokorski, *Heavy Flavor Multiplicities at Very High-Energies*, Nucl. Phys. B **155** (1979) 253–268.

[465] A. H. Mueller, *On the Multiplicity of Hadrons in QCD Jets*, Phys. Lett. B **104** (1981) 161–164.

[466] B. I. Ermolaev and V. S. Fadin, *Log - Log Asymptotic Form of Exclusive Cross-Sections in Quantum Chromodynamics*, JETP Lett. **33** (1981) 269–272.

[467] S. Catani, Y. L. Dokshitzer, F. Fiorani, and B. R. Webber, *Average number of jets in e+ e- annihilation*, Nucl. Phys. B **377** (1992) 445–460.

[468] I. M. Dremin and J. W. Gary, *Hadron multiplicities*, Phys. Rept. **349** (2001) 301–393, [hep-ph/0004215].

[469] **ATLAS** Collaboration, G. Aad et al., *Measurement of the charged-particle multiplicity inside jets from $\sqrt{s} = 8$ TeV pp collisions with the ATLAS detector*, Eur. Phys. J. C **76** (2016), no. 6 322, [arXiv:1602.00988].

[470] A. M. Polyakov, *A Similarity hypothesis in the strong interactions. 1. Multiple hadron production in e+ e- annihilation*, Zh. Eksp. Teor. Fiz. **59** (1970) 542–552.

[471] Z. Koba, H. B. Nielsen, and P. Olesen, *Scaling of multiplicity distributions in high-energy hadron collisions*, Nucl. Phys. B **40** (1972) 317–334.

[472] S. M. Stigler, *Stigler's law of eponymy**, Transactions of the New York Academy of Sciences **39** (1980), no. 1 Series II 147–157, [https://nyaspubs.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2164-0947.1980.tb02775.x].

[473] J. D. Bjorken, *Asymptotic Sum Rules at Infinite Momentum*, Phys. Rev. **179** (1969) 1547–1553.

[474] R. P. Feynman, *Very high-energy collisions of hadrons*, Phys. Rev. Lett. **23** (1969) 1415–1417.

[475] K. G. Wilson, *SOME EXPERIMENTS ON MULTIPLE PRODUCTION*, . CLNS-131.

[476] K. Konishi, A. Ukawa, and G. Veneziano, *Jet Calculus: A Simple Algorithm for Resolving QCD Jets*, Nucl. Phys. B **157** (1979) 45–107.

[477] Y. L. Dokshitzer, V. S. Fadin, and V. A. Khoze, *On the Sensitivity of the Inclusive Distributions in Parton Jets to the Coherence Effects in QCD Gluon Cascades*, Z. Phys. C **18** (1983) 37.

[478] E. D. Malaza and B. R. Webber, *QCD CORRECTIONS TO JET MULTIPLICITY MOMENTS*, Phys. Lett. B **149** (1984) 501–503.

[479] T. Munehisa and K. Tesima, *Integral Equation for Multiplicity Distribution*, Nucl. Phys. B **277** (1986) 575–592.

[480] Y. L. Dokshitzer, V. Khoze, A. Mueller, and S. Troyan, *Basics of perturbative qcd, ed, Frontieres, Gif-sur-Yvette France* (1991).

[481] F. Cuypers and K. Tesima, *Recoil effect on multiplicity correlation*, Z. Phys. C **54** (1992) 87–94.

[482] G. Gustafson, *Multiplicity distributions in QCD cascades*, Nucl. Phys. B **392** (1993) 251–280.

[483] Y. L. Dokshitzer, *Improved QCD treatment of the KNO phenomenon*, Phys. Lett. B **305** (1993) 295–301.

[484] G. Gustafson and M. Olsson, *Hadron and jet multiplicities in QCD cascades*, Nucl. Phys. B **406** (1993) 293–324.

[485] V. A. Khoze and W. Ochs, *Perturbative QCD approach to multiparticle production*, Int. J. Mod. Phys. A **12** (1997) 2949–3120, [`hep-ph/9701421`].

[486] **TASSO** Collaboration, M. Althoff et al., *Jet Production and Fragmentation in e+ e- Annihilation at 12-GeV to 43-GeV*, Z. Phys. C **22** (1984) 307–340.

[487] **L3** Collaboration, P. Achard et al., *Studies of hadronic event structure in e^+e^- annihilation from 30-GeV to 209-GeV with the L3 detector*, Phys. Rept. **399** (2004) 71–174, [`hep-ex/0406049`].

[488] **OPAL** Collaboration, K. Ackerstaff et al., *QCD studies with e+ e- annihilation data at 161-GeV*, Z. Phys. C **75** (1997) 193–207.

[489] K. Konishi, *MULTIPLICITY DISTRIBUTIONS IN QCD*, 5, 1979.

[490] F. Hayot and G. F. Sterman, *The Shape of the KNO Scaling Function in QCD*, Phys. Lett. B **121** (1983) 419–423.

[491] X. Cai and L.-S. Liu, *ON THE SCALING BEHAVIOR OF MULTIPLICITY DISTRIBUTION AT S**(1/2) = 540-GEV AND THE THREE FIREBALL MODEL*, Lett. Nuovo Cim. **37** (1983) 495–500.

[492] L.-s. Liu and T.-c. Meng, *Multiplicity and Energy Distributions in High-energy e^+e^-, pp and $p\bar{p}$ Collisions*, Phys. Rev. D **27** (1983) 2640.

[493] Y. L. Dokshitzer, V. S. Fadin, and V. A. Khoze, *Double Logs of Perturbative QCD for Parton Jets and Soft Hadron Spectra*, Z. Phys. C **15** (1982) 325.

[494] G. Pancheri and Y. N. Srivastava, *Quantum Chromodynamics Radiation and KNO Scaling*, Phys. Lett. B **128** (1983) 433–438.

[495] K.-c. Chou, L.-s. Liu, and T.-c. Meng, *KNO Scaling and Production Mechanism in High-energy Collisions*, Phys. Rev. D **28** (1983) 1080.

[496] P. Carruthers and C. C. Shih, *Correlations and Fluctuations in Hadronic Multiplicity Distributions: The Meaning of KNO Scaling*, Phys. Lett. B **127** (1983) 242–250.

[497] A. Bassetto, *KNO SCALING IN QCD JETS AND THE NEGATIVE BINOMIAL DISTRIBUTION*, Nucl. Phys. B **303** (1988) 703–712.

[498] Y. Liu, M. A. Nowak, and I. Zahed, *Mueller's dipole wave function in QCD: Emergent Koba-Nielsen-Olesen scaling in the double logarithm limit*, Phys. Rev. D **108** (2023), no. 3 034017, [`arXiv:2211.05169`].

[499] Y. Liu, M. A. Nowak, and I. Zahed, *Universality of Koba-Nielsen-Olesen scaling in QCD at high energy and entanglement*, `arXiv:2302.01380`.

[500] G. P. Korchemsky and G. F. Sterman, *Power corrections to event shapes and factorization*, Nucl. Phys. B **555** (1999) 335–351, [`hep-ph/9902341`].

[501] G. P. Korchemsky and S. Tafat, *On power corrections to the event shape distributions in QCD*, JHEP **10** (2000) 010, [`hep-ph/0007005`].

[502] S. W. Bosch, B. O. Lange, M. Neubert, and G. Paz, *Factorization and shape function effects in inclusive B meson decays*, Nucl. Phys. B **699** (2004) 335–386, [`hep-ph/0402094`].

[503] A. H. Hoang and I. W. Stewart, *Designing gapped soft functions for jet production*, Phys. Lett. B **660** (2008) 483–493, [`arXiv:0709.3519`].

[504] Z. Ligeti, I. W. Stewart, and F. J. Tackmann, *Treating the b quark distribution function with reliable uncertainties*, Phys. Rev. D **78** (2008) 114014, [`arXiv:0807.1926`].

[505] I. W. Stewart, F. J. Tackmann, and W. J. Waalewijn, *Dissecting Soft Radiation with Factorization*, Phys. Rev. Lett. **114** (2015), no. 9 092001, [`arXiv:1405.6722`].

[506] R. Akhoury and V. I. Zakharov, *On the universality of the leading, 1/Q power corrections in QCD*, Phys. Lett. B **357** (1995) 646–652, [`hep-ph/9504248`].

[507] Y. L. Dokshitzer and B. R. Webber, *Calculation of power corrections to hadronic event shapes*, Phys. Lett. B **352** (1995) 451–455, [`hep-ph/9504219`].

[508] C. Lee and G. F. Sterman, *Universality of nonperturbative effects in event shapes*, eConf **C0601121** (2006) A001, [`hep-ph/0603066`].

[509] C. Lee and G. F. Sterman, *Momentum Flow Correlations from Event Shapes: Factorized Soft Gluons and Soft-Collinear Effective Theory*, Phys. Rev. D **75** (2007) 014022, [`hep-ph/0611061`].

[510] M. Planck, *Sitzunsber: Deutsch, Akad. Wiss. Berlin* **33** (1923) 355.

[511] B. Decomps and A. Kastler, *Distribution of n particles between g cells. law of fluctuations*, *Compt. Rend.* **256** (1963).

[512] D. M. Blei, *Probabilistic topic models*, *Communications of the ACM* **55** (2012), no. 4 77–84.

[513] S. Arora, R. Ge, and A. Moitra, *Learning topic models–going beyond svd*, in *2012 IEEE 53rd annual symposium on foundations of computer science*, pp. 1–10, IEEE, 2012.

[514] B. Peon, *Is Hinchliffe's Rule True?*, *Annals of Gnosis* (8, 1988).

[515] **OPAL** Collaboration, G. Alexander et al., *A Direct observation of quark - gluon jet differences at LEP*, *Phys. Lett. B* **265** (1991) 462–474.

[516] **ALEPH** Collaboration, D. Buskulic et al., *Study of the subjet structure of quark and gluon jets*, *Phys. Lett. B* **346** (1995) 389–398.

[517] **OPAL** Collaboration, R. Akers et al., *A Model independent measurement of quark and gluon jet properties and differences*, *Z. Phys. C* **68** (1995) 179–202.

[518] **DELPHI** Collaboration, P. Abreu et al., *Energy dependence of the differences between the quark and gluon jet fragmentation*, *Z. Phys. C* **70** (1996) 179–196.

[519] **OPAL** Collaboration, K. Ackerstaff et al., *Multiplicity distributions of gluon and quark jets and tests of QCD analytic predictions*, *Eur. Phys. J. C* **1** (1998) 479–494, [`hep-ex/9708029`].

[520] **ATLAS** Collaboration, G. Aad et al., *Performance and calibration of quark/gluon-jet taggers using 140 fb$^{?1}$ of pp collisions at TeV with the ATLAS detector**, *Chin. Phys. C* **48** (2024), no. 2 023001, [`arXiv:2308.00716`].

[521] R. Abdul Khalek et al., *Science Requirements and Detector Concepts for the Electron-Ion Collider: EIC Yellow Report*, *Nucl. Phys. A* **1026** (2022) 122447, [`arXiv:2103.05419`].

[522] A. Einstein, B. Podolsky, and N. Rosen, *Can quantum-mechanical description of physical reality be considered complete?*, *Physical review* **47** (1935), no. 10 777.

[523] D. Bohm, *A suggested interpretation of the quantum theory in terms of" hidden" variables. i*, *Physical review* **85** (1952), no. 2 166.

[524] J. S. Bell, *On the Einstein-Podolsky-Rosen paradox*, *Physics Physique Fizika* **1** (1964) 195–200.

[525] J. S. Bell, *On the problem of hidden variables in quantum mechanics*, *Reviews of Modern physics* **38** (1966), no. 3 447.

[526] Y. Afik and J. R. M. n. de Nova, *Entanglement and quantum tomography with top quarks at the LHC*, *Eur. Phys. J. Plus* **136** (2021), no. 9 907, [`arXiv:2003.02280`].

[527] M. Fabbrichesi, R. Floreanini, and G. Panizzo, *Testing Bell Inequalities at the LHC with Top-Quark Pairs*, *Phys. Rev. Lett.* **127** (2021), no. 16 161801, [`arXiv:2102.11883`].

[528] Y. Takubo, T. Ichikawa, S. Higashino, Y. Mori, K. Nagano, and I. Tsutsui, *Feasibility of Bell inequality violation at the ATLAS experiment with flavor entanglement of B0B̄0 pairs from pp collisions*, *Phys. Rev. D* **104** (2021), no. 5 056004, [`arXiv:2106.07399`].

[529] A. J. Barr, *Testing Bell inequalities in Higgs boson decays*, *Phys. Lett. B* **825** (2022) 136866, [`arXiv:2106.01377`].

[530] C. Severi, C. D. E. Boschi, F. Maltoni, and M. Sioli, *Quantum tops at the LHC: from entanglement to Bell inequalities*, *Eur. Phys. J. C* **82** (2022), no. 4 285, [`arXiv:2110.10112`].

[531] Y. Afik and J. R. M. n. de Nova, *Quantum information with top quarks in QCD*, *Quantum* **6** (2022) 820, [`arXiv:2203.05582`].

[532] **ATLAS** Collaboration, G. Aad et al., *Observation of quantum entanglement in top-quark pairs using the ATLAS detector*, `arXiv:2311.07288`.

[533] **CMS** Collaboration, T. C. Collaboration, *Observation of quantum entanglement in top quark pair production in proton-proton collisions at $\sqrt{s}$ = 13 TeV*, `arXiv:2406.03976`.

[534] D. M. Greenberger, M. A. Horne, A. Shimony, and A. Zeilinger, *Bell's theorem without inequalities*, *American Journal of Physics* **58** (1990), no. 12 1131–1143.

[535] N. D. Mermin, *Quantum mysteries revisited*, *Am. J. Phys* **58** (1990), no. 8 731–734.

[536] N. D. Mermin, *What's wrong with these elements of reality?*, *Physics Today* **43** (1990), no. 6 9–11.

[537] S. Coleman, *Sidney Coleman's Dirac Lecture "Quantum Mechanics in Your Face"*, `arXiv:2011.12671`. `https://www.youtube.com/watch?v=EtyNMlXN-sw`.

[538] Z. Dong, D. Gonçalves, K. Kong, and A. Navarro, *When the Machine Chimes the Bell: Entanglement and Bell Inequalities with Boosted $t\bar{t}$*, `arXiv:2305.07075`.

[539] T. Han, M. Low, and T. A. Wu, *Quantum Entanglement and Bell Inequality Violation in Semi-Leptonic Top Decays*, `arXiv:2310.17696`.

[540] F. Maltoni, C. Severi, S. Tentori, and E. Vryonidou, *Quantum tops at circular lepton colliders*, `arXiv:2404.08049`.

[541] B. Tweedie, *Better Hadronic Top Quark Polarimetry*, *Phys. Rev. D* **90** (2014), no. 9 094010, [`arXiv:1401.3021`].

[542] J. Duarte-Campderros, G. Perez, M. Schlaffer, and A. Soffer, *Probing the Higgs–strange-quark coupling at e^+e^- colliders using light-jet flavor tagging*, *Phys. Rev. D* **101** (2020), no. 11 115005, [`arXiv:1811.09636`].

[543] Y. Nakai, D. Shih, and S. Thomas, *Strange Jet Tagging*, `arXiv:2003.09517`.

[544] A. Albert et al., *Strange quark as a probe for new physics in the Higgs sector*, in *Snowmass 2021*, 3, 2022. `arXiv:2203.07535`.

[545] F. Bedeschi, L. Gouskos, and M. Selvaggi, *Jet flavour tagging for future colliders with fast simulation*, *Eur. Phys. J. C* **82** (2022), no. 7 646, [`arXiv:2202.03285`].

[546] J. Erdmann, O. Nackenhorst, and S. V. Zeißner, *Maximum performance of strange-jet tagging at hadron colliders*, *JINST* **16** (2021), no. 08 P08039, [`arXiv:2011.10736`].

[547] R. D. Field and R. P. Feynman, *A Parametrization of the Properties of Quark Jets*, *Nucl. Phys. B* **136** (1978) 1.

[548] **Fermilab-Serpukhov-Moscow-Michigan** Collaboration, J. P. Berge et al., *Net Charge in Deep Inelastic Antineutrino - Nucleon Scattering*, *Phys. Lett. B* **91** (1980) 311–313.

[549] R. Erickson et al., *Charge Retention in Deep Inelastic Electroproduction*, *Phys. Rev. Lett.* **42** (1979) 822. [Erratum: Phys.Rev.Lett. 42, 1246 (1979)].

[550] J. P. Berge et al., *Quark Jets from Antineutrino Interactions 1: Net Charge and Factorization in the Quark Jets*, *Nucl. Phys. B* **184** (1981) 13–30.

[551] **Aachen-Bonn-CERN-Munich-Oxford** Collaboration, P. Allen et al., *Multiplicity Distributions in Neutrino - Hydrogen Interactions*, *Nucl. Phys. B* **181** (1981) 385–402.

[552] **Aachen-Bonn-CERN-Munich-Oxford** Collaboration, P. Allen et al., *Charge Properties of the Hadronic System in Neutrino p and Anti-neutrino p Interactions*, *Phys. Lett. B* **112** (1982) 88.

[553] **European Muon** Collaboration, J. P. Albanese et al., *Quark Charge Retention in Final State Hadrons From Deep Inelastic Muon Scattering*, *Phys. Lett. B* **144** (1984) 302–308.

[554] **Amsterdam-Bologna-Padua-Pisa-Saclay-Turin** Collaboration, S. Barlag et al., *Charged Hadron Multiplicities in High-energy Anti-muon Neutrino n and Anti-muon Neutrino p Interactions*, *Z. Phys. C* **11** (1982) 283. [Erratum: Z.Phys.C 14, 281 (1982)].

[555] **TASSO** Collaboration, W. Braunschweig et al., *Measurement of the Asymmetry of b Quark Production in e^+e^- Annihilation of $\sqrt{s} = 35$-GeV*, *Z. Phys. C* **48** (1990) 433–444.

[556] **DELPHI** Collaboration, P. Abreu et al., *A Measurement of sin**2 theta(W) from the charge asymmetry of hadronic events at the Z0 peak*, *Phys. Lett. B* **277** (1992) 371–382.

[557] **ALEPH** Collaboration, D. Decamp et al., *Measurement of charge asymmetry in hadronic Z decays*, *Phys. Lett. B* **259** (1991) 377–388.

[558] **OPAL** Collaboration, P. D. Acton et al., *A Measurement of the forward - backward charge asymmetry in hadronic decays of the Z0*, *Phys. Lett. B* **294** (1992) 436–450.

[559] **OPAL** Collaboration, R. Akers et al., *Measurement of the time dependence of B(d)0 <—> anti-B(d)0 mixing using a jet charge technique*, *Phys. Lett. B* **327** (1994) 411–424.

[560] **SLD** Collaboration, K. Abe et al., *Measurement of A(b) from the left-right forward - backward asymmetry of b quark production in Z0 decays using a momentum weighted track charge technique*, *Phys. Rev. Lett.* **74** (1995) 2890–2894.

[561] **DELPHI** Collaboration, P. Abreu et al., *Measurement of the B(d)0 oscillation frequency using kaons, leptons and jet charge*, *Z. Phys. C* **72** (1996) 17–30.

[562] **ALEPH** Collaboration, R. Barate et al., *Measurement of triple gauge boson couplings at 172-GeV*, *Phys. Lett. B* **422** (1998) 369–383.

[563] **CDF** Collaboration, F. Abe et al., *Measurement of $B^0\overline{B}^0$ flavor oscillations using jet-charge and lepton flavor tagging in $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV*, *Phys. Rev. D* **60** (1999) 072003, [hep-ex/9903011].

[564] **L3** Collaboration, M. Acciarri et al., *Measurement of triple gauge boson couplings of the W boson at LEP*, *Phys. Lett. B* **467** (1999) 171–184, [hep-ex/9910008].

[565] **OPAL** Collaboration, G. Abbiendi et al., *Measurement of triple gauge boson couplings from W^+W^- production at LEP energies up to 189-GeV*, *Eur. Phys. J. C* **19** (2001) 1–14, [hep-ex/0009022].

[566] **DELPHI** Collaboration, P. Abreu et al., *Measurement of trilinear gauge boson couplings WWV, (V ≡ Z, γ) in e⁺e⁻ collisions at 189-GeV, Phys. Lett. B* **502** (2001) 9–23, [`hep-ex/0102041`].

[567] **D0** Collaboration, V. M. Abazov et al., *Experimental discrimination between charge 2e/3 top quark and charge 4e/3 exotic quark production scenarios, Phys. Rev. Lett.* **98** (2007) 041801, [`hep-ex/0608044`].

[568] **CDF** Collaboration, T. Aaltonen et al., *Exclusion of exotic top-like quarks with -4/3 electric charge using jet-charge tagging in single-lepton ttbar events at CDF, Phys. Rev. D* **88** (2013), no. 3 032003, [`arXiv:1304.4141`].

[569] D. Krohn, M. D. Schwartz, T. Lin, and W. J. Waalewijn, *Jet Charge at the LHC, Phys. Rev. Lett.* **110** (2013), no. 21 212001, [`arXiv:1209.2421`].

[570] W. J. Waalewijn, *Calculating the Charge of a Jet, Phys. Rev. D* **86** (2012) 094030, [`arXiv:1209.3019`].

[571] **ATLAS** Collaboration, G. Aad et al., *Measurement of the top quark charge in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector, JHEP* **11** (2013) 031, [`arXiv:1307.4568`].

[572] **CMS** Collaboration, V. Khachatryan et al., *Identification techniques for highly boosted W bosons that decay into hadrons, JHEP* **12** (2014) 017, [`arXiv:1410.4227`].

[573] **ATLAS** Collaboration, G. Aad et al., *Measurement of jet charge in dijet events from $\sqrt{s}$=8 TeV pp collisions with the ATLAS detector, Phys. Rev. D* **93** (2016), no. 5 052003, [`arXiv:1509.05190`].

[574] **CMS** Collaboration, A. M. Sirunyan et al., *Measurements of jet charge with dijet events in pp collisions at $\sqrt{s} = 8$ TeV, JHEP* **10** (2017) 131, [`arXiv:1706.05868`].

[575] **CMS** Collaboration, A. M. Sirunyan et al., *Measurement of quark- and gluon-like jet fractions using jet charge in PbPb and pp collisions at 5.02 TeV, JHEP* **07** (2020) 115, [`arXiv:2004.00602`].

[576] Z.-B. Kang, A. J. Larkoski, and J. Yang, *Towards a Nonperturbative Formulation of the Jet Charge, Phys. Rev. Lett.* **130** (2023), no. 15 151901, [`arXiv:2301.09649`].

[577] T.-J. Stieltjes, *Recherches sur les fractions continues*, in *Annales de la Faculté des sciences de Toulouse: Mathématiques*, vol. 8, pp. J1–J122, 1894.

[578] H. Hamburger, *Über eine erweiterung des stieltjesschen momentenproblems, Mathematische Annalen* **81** (1920), no. 2-4 235–319.

[579] F. Hausdorff, *Summationsmethoden und momentfolgen. i, Mathematische Zeitschrift* **9** (1921), no. 1-2 74–109.

[580] T. Carleman, *Ueber die approximation analytischer funktionen durch lineare aggregate von vorgegebenen potenzen, Arkiv för mat., astron. o. fys* **17** (1922) 8.

[581] M. Krein, *On a problem of extrapolation of an kolmogorov*, in *Dokl. Akad. Nauk SSSR*, vol. 46, p. 376, 1945.

[582] M. B. Green and C. Wen, *Superstring amplitudes, unitarily, and Hankel determinants of multiple zeta values, JHEP* **11** (2019) 079, [`arXiv:1908.08426`].

[583] W.-M. Chen, Y.-T. Huang, T. Noumi, and C. Wen, *Unitarity bounds on charged/neutral state mass ratios, Phys. Rev. D* **100** (2019), no. 2 025016, [`arXiv:1901.11480`].

[584] N. Arkani-Hamed, Y.-T. Huang, and S.-H. Shao, *On the Positive Geometry of Conformal Field Theory*, JHEP **06** (2019) 124, [arXiv:1812.07739].

[585] K. Sen, A. Sinha, and A. Zahed, *Positive geometry in the diagonal limit of the conformal bootstrap*, JHEP **11** (2019) 059, [arXiv:1906.07202].

[586] N. Arkani-Hamed, T.-C. Huang, and Y.-t. Huang, *The EFT-Hedron*, JHEP **05** (2021) 259, [arXiv:2012.15849].

[587] B. Bellazzini, J. Elias Miró, R. Rattazzi, M. Riembau, and F. Riva, *Positive moments for scattering amplitudes*, Phys. Rev. D **104** (2021), no. 3 036006, [arXiv:2011.00037].

[588] Y.-t. Huang, J.-Y. Liu, L. Rodina, and Y. Wang, *Carving out the Space of Open-String S-matrix*, JHEP **04** (2021) 195, [arXiv:2008.02293].

[589] A. J. Tolley, Z.-Y. Wang, and S.-Y. Zhou, *New positivity bounds from full crossing symmetry*, JHEP **05** (2021) 255, [arXiv:2011.02400].

[590] Z. Bern, D. Kosmopoulos, and A. Zhiboedov, *Gravitational effective field theory islands, low-spin dominance, and the four-graviton amplitude*, J. Phys. A **54** (2021), no. 34 344002, [arXiv:2103.12728].

[591] Z.-B. Kang, R. Kao, and A. J. Larkoski, *Multiplicity scaling of fragmentation function*, Phys. Rev. D **109** (2024), no. 5 054039, [arXiv:2305.13359].

[592] R. Bhatia and C. Davis, *A better bound on the variance*, The American Mathematical Monthly **107** (2000), no. 4 353–357, [https://doi.org/10.1080/00029890.2000.12005203].

[593] K. Fraser and M. D. Schwartz, *Jet Charge and Machine Learning*, JHEP **10** (2018) 093, [arXiv:1803.08066].

[594] Y.-C. J. Chen, C.-W. Chiang, G. Cottin, and D. Shih, *Boosted W and Z tagging with jet charge and deep learning*, Phys. Rev. D **101** (2020), no. 5 053001, [arXiv:1908.08256].

[595] Z.-B. Kang, X. Liu, S. Mantry, M. C. Spraker, and T. Wilson, *Dynamic Jet Charge*, Phys. Rev. D **103** (2021), no. 7 074028, [arXiv:2101.04304].

[596] K. Lee, J. Mulligan, M. Płoskoń, F. Ringer, and F. Yuan, *Machine learning-based jet and event classification at the Electron-Ion Collider with applications to hadron structure and spin physics*, JHEP **03** (2023) 085, [arXiv:2210.06450].

[597] **DELPHI** Collaboration, P. Abreu et al., *Measurement of the quark and gluon fragmentation functions in Z0 hadronic decays*, Eur. Phys. J. C **6** (1999) 19–33.

[598] **OPAL** Collaboration, G. Abbiendi et al., *Scaling violations of quark and gluon jet fragmentation functions in e+ e- annihilations at s**(1/2) = 91.2-GeV and 183-GeV to 209-GeV*, Eur. Phys. J. C **37** (2004), no. 1 25–47, [hep-ex/0404026].

[599] **CMS** Collaboration, S. Chatrchyan et al., *Search for a Higgs Boson in the Decay Channel $H \to ZZ^* \to q\bar{q}\ell^-\ell^+$ in pp Collisions at $\sqrt{s} = 7$ TeV*, JHEP **04** (2012) 036, [arXiv:1202.1416].

[600] **CMS** Collaboration, *Pileup Jet Identification*, 2013. CMS-PAS-JME-13-005.

[601] A. J. Larkoski, J. Thaler, and W. J. Waalewijn, *Gaining (Mutual) Information about Quark/Gluon Discrimination*, JHEP **11** (2014) 129, [arXiv:1408.3122].