Jet Physics Lecture 2

HCPSS 2020 Andrew Larkoski

Welcome back to the second lecture on jets and machine learning. In this lecture, we're going to think like a machine to understand the problem of binary discrimination, or, techniques for distinguishing two samples mixed in an ensemble. Before we get to that, I want to wrap up a couple points from the previous lecture that we will use to great effect this lecture. The first is the property of infrared and collinear safety of the angularities. We had found that the Sudakov form factor for the angularity measured on a quark jet took the form:

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

for $\tau_\alpha \in [0,1]$ and $\alpha > 0$. Recall that α was the angular exponent of the angularity:

$$\tau_\alpha = \sum_{i\in J} z_i \theta_i^\alpha,$$ and I said that it must be positive

for infrared and collinear safety. Using our axiom of scale invariance of QCD, we were lead to a probability for single gluon emission that diverged in the soft (=low-energy) and collinear limits. When all degenerate configurations of emitted gluons are summed up, we generate the Sudakov factor, which is finite, by the KLN theorem. However, the

Sudakov factor is a series with terms to <u>all</u> orders in the coupling α_s. What if we just want a description of QCD and jets to a <u>fixed</u> order in α_s? That is, a description as provided by some collection of Feynman diagrams? Well, for the result to be sensible (i.e., finite) even though the fundamental probability distribution diverges in the soft and/or collinear limits requires a delicate property of the observables that we choose to measure on the jet. In particular, for the distribution of an observable to be finite <u>almost everywhere</u> in its domain requires that the soft and collinear limits of that observable map to a unique value. That is, there is a <u>single</u> value of the observable for which <u>all</u> divergences from soft or collinear gluon emission are located. Another way to state this criteria is that exactly 0 energy or exactly collinear gluons do <u>not</u> affect the value of the observable. Such an observable for which this is true is called "infrared and collinear safe" or IRC safe. Isolating all divergences to a single value of the observable means that away from that value, everything is well-defined and finite.

The angularities are IRC safe because soft

and collinear gluons do not contribute to τ_α, for $\alpha > 0$. However, not all possible observables or questions you can ask of a jet are IRC safe. Perhaps the canonical example of non-IRC safety is that of <u>multiplicity</u>, or number of particles in a jet. A jet could consist of a single, bare quark so multiplicity would be 1. However, say that quark emits an exactly collinear gluon; now multiplicity would be two. However, this exactly collinear emission is degenerate to the bare quark, and violates the assumptions of KLN for finiteness. So, multiplicity is <u>not</u> IRC safe.

Another thing to address now is what the Sudakov form factor is for a high-energy gluon jet. From what we have discussed thus far, the only difference between a quark and a gluon is the color that either carry. (Implicitly also their spin is importantly different, but we won't pursue that more here.) A quark is in the fundamental representation of SU(3) color, and so the number of colors it can share with a soft/collinear gluon is controlled by the fundamental quadratic Casimir, $C_F = 4/3$. Gluons, by contrast live in the adjoint representation of SU(3) color, and can share <u>more</u> color with soft/collinear gluons than can quarks. The adjoint quadratic Casimir $C_A = 3$ in

QCD, so this has consequences for the gluon Sudakov factor and a heuristic for understanding properties of QCD jets. The gluon Sudakov is found by replacing $C_F \to C_A$ in the quark Sudakov:

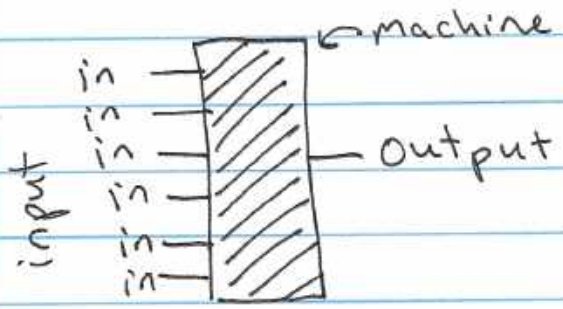
$$\Sigma_g(\tau_\alpha) = \exp\left[-\frac{\alpha_s}{\pi}\frac{C_A}{\alpha}\ln^2\tau_\alpha\right].$$

Because $C_A > C_F$, gluons are more likely to emit soft/collinear gluons, so it is more difficult for a gluon jet to have a very small value of τ_α. Hence, the Sudakov provides greater exponential suppression for gluons than quarks.

With those points established, I now want to pivot to discussing <u>machine learning</u>, in a very restricted (and biased!) manner. Machine learning is exploding as a discipline in particle physics, so there is no hope for me to discuss broadly how it is employed. However, my biased viewpoint is from that of a theorist who is selfishly interested in learning more about nature. So how can we use machine learning to learn more as flesh-and-blood humans? Let's first define what we are working with.

My theorist definition of a machine (or neural network or any fancy computer science algorithm) is the

following. A <u>machine</u> is a black box which takes <u>input</u> and returns <u>output</u>:

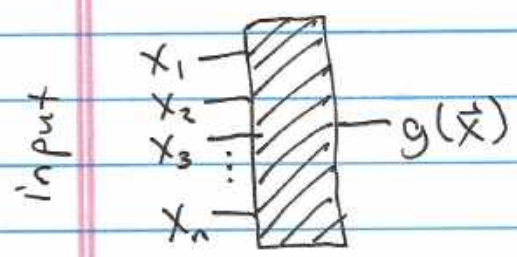


By "black box" I mean that the machine performs some manipulation on the input to produce the output, but the way in which it does it is unknowable to me. Don't worry; it will turn out fine that we don't know what the machine does. I have also illustrated the input as multiple entries and the output as a single result. That is, we consider the input to be some n dimensional vector $\vec{x}$ and the output to be a single number, g. So, all the machine is is a <u>function</u> of the input:

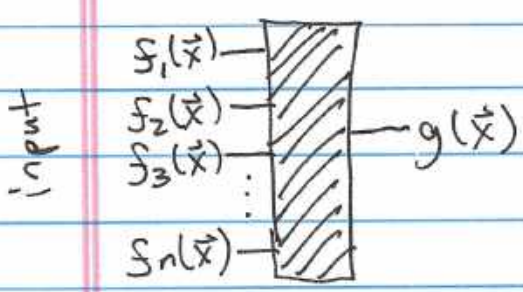
Machine: $\mathbb{R}^n \to \mathbb{R}$, which can be represented

by the function $g \equiv g(\vec{x})$. One can consider more general inputs and outputs, but we will keep it restricted to this scalar function case, again for simplicity.

The first result that allows us to learn anything at all is the <u>universal approximation theorem</u>. For our purposes, the statement of the universal approximation theorem is the following: A sufficiently large and powerful machine can output an <u>arbitrary</u> function of the inputs. Concretely, let's say that the input data are the entries of the vector $\vec{x}$, and the output of the machine is the function $g(\vec{x})$:



The universal approximation theorem states that we can modify the inputs to (almost) any collection of functions $\{f_i(\vec{x})\}_{i=1}^{n}$ and the machine will return (or "learn") the same function $g(\vec{x})$:



As long as the collection of f_i functions contains the same <u>total</u> information as the $\vec{x}$ input data, the machine will always learn $g(\vec{x})$. So, the universal approximation theorem states that we can choose a convenient form of input data, and so we can optimize for a form we understand well.

Now, a lot of machine learning practitioners will say that the universal approximation theorem is less than unuseful, because the actual behavior of the machine depends so sensitively on precise implementation and architecture. However, I again want to emphasize that we are <u>not</u> interested here in determining how a particular machine performs; we want to <u>think</u> like a machine to learn something about Nature.

The specific task we would like the machine to perform is <u>binary discrimination</u>, or, signal vs. background separation. The formulation of a binary discrimination problem is that we want the machine to separate, as efficiently as possible, signal from background events in a mixed ensemble. What we mean by "signal" and "background" is that signal events, or a collection of the input data, are drawn from the probability distribution $p_s(\vec{x})$, while background events are drawn from the probability distribution $p_b(\vec{x})$. Given a collection of identified signal and background events, the machine learns what the probability distributions are, and correspodingly outputs the probability with which it believes that an individual event is signal-type or background-type.

It doesn't do this separation blindly; we know the <u>optimal</u> binary discriminant from the result of the Neyman-Pearson Lemma. In the 1930s, Jerzy Neyman and Egon Pearson proved that the <u>likelihood ratio</u> is the optimal binary discriminant. The likelihood ratio $\mathcal{L}$ is simply the ratio of signal to background probability distributions:

$$\mathcal{L} = \frac{P_s(\vec{x})}{P_b(\vec{x})}.$$

The likelihood naturally and beautifully separates signal and background, to the maximal extent provided by the probability distributions. If $\mathcal{L} \to 0$, then the background probability is large compared to signal, and vice-versa for $\mathcal{L} \to \infty$. Thus:

$\mathcal{L} \to 0 \Rightarrow$ pure background
$\mathcal{L} \to \infty \Rightarrow$ pure signal

Additionally, a larger class of quantities than strictly just the likelihood are equal in discrimination power. <u>Any</u> function monotonic in the likelihood $\mathcal{L}$ is equivalent in discrimination power. This monotonic function can be exploited to simplify the range of values that the discrimination observable assumes. For example, the function

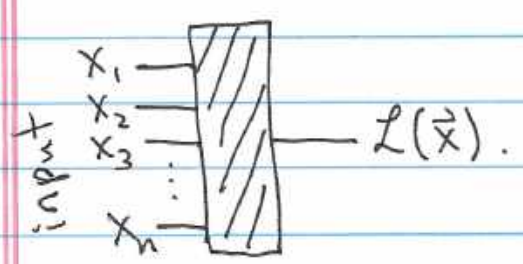
$$h(\mathcal{L}) = \frac{\mathcal{L}}{1+\mathcal{L}} = \frac{p_s(\vec{x})}{p_s(\vec{x}) + p_b(\vec{x})},$$

is monotonic in $\mathcal{L} \in [0,\infty)$, but nicely maps to the domain $h(\mathcal{L}) \in [0,1]$. In particular:

$h(\mathcal{L} \to \infty) \to 1$, signal-pure,
$h(\mathcal{L} \to 0) \to 0$, background-pure.

Again, there's nothing special about this $h(\mathcal{L})$, but it can be a nice object to consider.

So, for the case at hand, our machine takes in a collection of input data $\vec{x}$ for signal and background events drawn from probability distributions $p_s(\vec{x})$ and $p_b(\vec{x})$, respectively, and outputs the likelihood ratio, or a monotonic function of it:



With that set-up, let's now describe the physical system we would like to learn about. Our goal in the rest of this lecture is to learn something about the likelihood ratio for the problem of discrimination of jets initiated by high-energy quarks versus those jets

initiated by high-energy gluons. The input data that we will use in our machine is the collection of four-momenta of particles in quark or gluon jets. Let's call this set of momenta $\{p_i\}_{i\in J}$, and the quark and gluon probability distributions are thus:

$$P_q(\{p_i\}_{i\in J}), \quad P_g(\{p_i\}_{i\in J}),$$

where J is the jet of interest (i.e., the collection of soft and collimated particles). The likelihood ratio is then

$$\mathcal{L} = \frac{P_g(\{p_i\}_{i\in J})}{P_q(\{p_i\}_{i\in J})}.$$

So, there we go. We're done, right? While true, there is almost zero information in these statements; specifically, we know essentially nothing about the functional form of $\mathcal{L}$ for quark and gluon discrimination. Can we learn <u>any</u> information that will help us in our task?

The first problem with this formulation is the explicit use of particle four-vectors as the input data. This is a rather poor way to organize the information in a jet because, by the soft and collinear singularities of QCD, a substantial amount of particles will have (nearly) degenerate momenta, which is challenging to

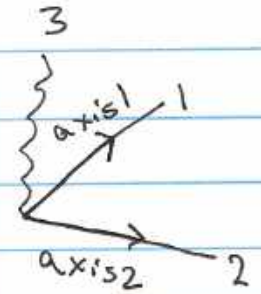
interpret theoretically so we humans can actually learn something. So, using the universal approximation theorem, let's see if we can reorganize the information contained in four-vectors into a much more useful, and human-interpretable, form.

First, let's see what we are dealing with and what the dimension of the input space actually is. A generic four-vector has four real components, so the four-vectors of N particles is (naively) some 4N real dimensional space. However, as <u>real</u> particles, their momenta are all on-shell. For simplicity (but no other reason) let's assume that all particles are massless. As such, an on-shell, massless four-vector actually only has 3 degrees of freedom, so the space of input momenta is only 3N real dimensions. Additionally, total energy and momentum are conserved, which imposes 4 further linear constraints on all momenta. Thus, there are 3N-4 degrees of freedom to completely define the four momenta of N particles in a jet (assuming on-shellness and total momentum conservation). So, we just need to identify 3N-4 functions of the particles' momenta appropriately and we can use the universal approximation theorem to claim that our machine would find the <u>same</u> likelihood ratio.

So, what functions of momenta should we use? This is really a matter of taste, but if we want to exploit our perturbative understanding of QCD to the problem at hand, then we would want these functions to be infrared and collinear safe observables. What 3N-4 IRC safe observables should we use? This is much more a matter of taste, and there are many possible answers, but here we will just consider the N-subjettiness observables. N-subjettiness is a class of IRC safe observables that extends the angularities to resolve N prongs in a jet. As a function of energy fractions z_i and angles, N-subjettiness $\tau_N^{(\alpha)}$ is defined to be:

$$\tau_N^{(\alpha)} = \sum_{i \in J} z_i \min\{\theta_{i1}^\alpha, \theta_{i2}^\alpha, \ldots, \theta_{iN}^\alpha\}, \text{ and } \alpha > 0.$$

Here θ_{iK} is the angle between particle i's momentum and axis K in the jet. Defining N-subjettiness requires placing N axes in the jet, nominally in the directions of dominant energy flow. For example, consider 2-subjettiness measured on a jet with two hard particles, 1 and 2, and one soft particle 3: The two axes would, for example, align with particles 1 and 2 and only particle 3 would contribute to $\tau_2^{(\alpha)}$!



$$\Rightarrow \tau_2^{(\alpha)} = \sum_{i \in J} z_i \min\{\theta_{i1}^{\alpha}, \theta_{i2}^{\alpha}\}$$

$$= z_3 \min\{\theta_{31}^{\alpha}, \theta_{32}^{\alpha}\}$$

Angularities are 1-subjettiness, and IRC safety of N-subjettinesses essentially follows from IRC safety of angularities (with some caveats regarding axes). N-subjettinesses have the added benefit that they are <u>additive</u>, and so multiple soft and collinear emissions in the jet generates a Sudakov factor, exactly as we observed for angularities.

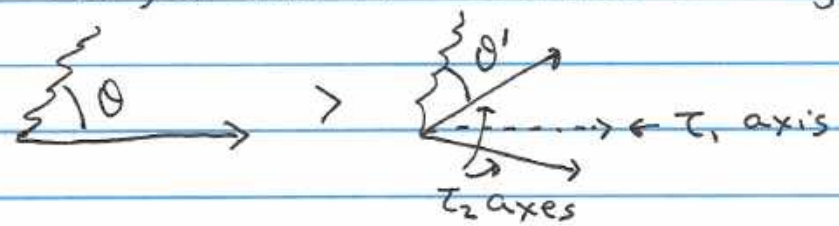
So, we already know a lot about the space of N-subjettinesses, so we can choose 3N-4 of them to resolve the four-momenta of N particles in the jet. Because time is short, I won't go into explicit details about what that collection of N-subjettinesses should be to ensure that they have the same information as the collection of four-vectors. See arXiv: 1704.08249 for all the details.

Now we're cooking. What properties of the likelihood ratio $\mathcal{L}$ for quark vs. gluon discrimination can we learn using the N-subjettiness variables as inputs to our machine (which in our case is our brains!)?

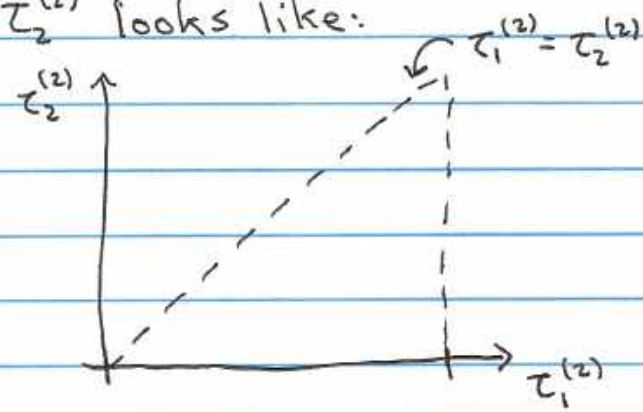
I'll present a simplified argument here, and more details can be found in 1906.01639. Let's just imagine, for simplicity, our entire input space was just that of $\tau_1^{(2)}$ and $\tau_2^{(2)}$, the one- and two-subjettiness with angular exponents equal to 2:

$$\tau_1^{(2)} = \sum_{i \in J} z_i \theta_i^2 \,, \quad \tau_2^{(2)} = \sum_{i \in J} z_i \min\{\theta_{i1}^2, \theta_{i2}^2\}.$$

Note that both $\tau_1^{(2)}, \tau_2^{(2)} > 0$ and $\tau_1^{(2)} > \tau_2^{(2)}$ because with two axes in the jet the distance of any particle to those axes is less than or equal to the distance to a single axis in the center of the jet:



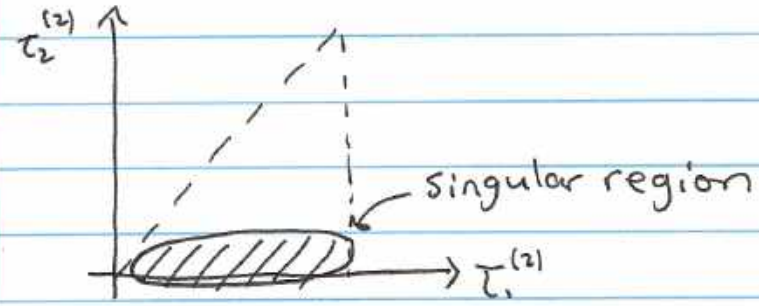
Thus, our phase space defined by measuring $\tau_1^{(2)}$ and $\tau_2^{(2)}$ looks like:



Phase space is the triangular region bounded by the τ_1 axis and the dashed lines.

To go further we need two things: (1) identification of the soft/collinear region and (2) the form

of the likelihood on this space. (1) is easy enough to answer: because $\tau_N^{(\alpha)} > 0$ and is IRC safe, the $\tau_N^{(\alpha)} \to 0$ limit is the soft/collinear divergent limit. On our phase space, this is just the region near the $\tau_1^{(2)}$ axis:



Okay, let's see <u>why</u> this observation is so important. The likelihood ratio of quark and gluon probability distributions on this space is:

$$\mathcal{L}(\tau_1^{(2)}, \tau_2^{(2)}) = \frac{P_g(\tau_1^{(2)}, \tau_2^{(2)})}{P_q(\tau_1^{(2)}, \tau_2^{(2)})}$$

Do we have any information as to what the functional form of this ratio might be? Yes, from the Sudakovs! By the IRC safety (and additivity) of N-subjettiness, the quark and gluon probabilities have Sudakov factors of the form:

$$P_q(\tau_1^{(2)}, \tau_2^{(2)}) \sim e^{-\alpha_s C_F f(\tau_1^{(2)}, \tau_2^{(2)})}, \quad P_g(\tau_1^{(2)}, \tau_2^{(2)}) \sim e^{-\alpha_s C_A f(\tau_1^{(2)}, \tau_2^{(2)})}.$$

Now, without an explicit calculation, we don't know what the functional form of $f(\tau_1^{(2)}, \tau_2^{(2)})$ is, but, by IRC safety of N-Subjettiness, we know its behavior in limits. The soft/collinear limit corresponds to $\tau_2^{(2)} \to 0$ (with $\tau_1^{(2)} > \tau_2^{(2)}$), and in this limit, the Sudakov factor exponentially suppresses the probability distributions. With this exponential form, we then must have that

$$f(\tau_1^{(2)}, \tau_2^{(2)} \to 0) \to \infty .$$

Now, the form of the likelihood ratio for quark versus gluon jet discrimination is

$$\mathcal{L} = \frac{P_g(\tau_1^{(2)}, \tau_2^{(2)})}{P_q(\tau_1^{(2)}, \tau_2^{(2)})} \sim e^{-\alpha_s (C_A - C_F) f(\tau_1^{(1)}, \tau_2^{(2)})}$$

Because $C_A > C_F$ in QCD, we <u>still</u> have that the likelihood vanishes in the singular, $\tau_2^{(2)} \to 0$ limit:

$$\mathcal{L}(\tau_1^{(2)}, \tau_2^{(2)}) \to 0 \quad \text{as } \tau_2^{(2)} \to 0 .$$

<u>However</u>, the <u>entire</u> region where $\tau_2^{(2)} \to 0$ is the soft/collinear limit in which fixed-order description of jets diverges. In this entire singular region, the likelihood takes a unique value: $\mathcal{L} = 0$. Thus, the likelihood for quark vs. gluon discrimination is <u>IRC safe</u>,

from our earlier discussion. Out of <u>all</u> possible functions of input particle momenta, the likelihood for this problem is IRC safe, which is a <u>strong</u> constraint on its form. Thus, we learn that if you want to distinguish quark flavor from gluon flavor jets, a good place to start is to use an IRC safe observable.

Again, I want to emphasize that we, humans, learned something about QCD by thinking like a machine. What else might we learn with this approach? I hope you can find something new!

<u>Exercises</u>

(1) Just consider the angularities for discrimination of quark vs. gluon jets.

(a) What is the likelihood ratio for the probability distributions $p_g(\tau_a)$, $p_q(\tau_a)$?

(b) What is the distribution of this likelihood $\mathcal{L}$ on quark and gluon jets, $p_q(\mathcal{L})$ and $p_g(\mathcal{L})$?

(c) The <u>receiver</u> <u>operating</u> <u>characteristic</u> curve (ROC) quantifies the "strength" of separation power

of the likelihood, as a function of a cut on the likelihood. If the cumulative distributions of the likelihood for quarks and gluons are

$$\Sigma_q(\mathcal{L}) = \int_0^{\mathcal{L}} d\mathcal{L}' \, p_q(\mathcal{L}'), \quad \Sigma_g(\mathcal{L}) = \int_0^{\mathcal{L}} d\mathcal{L}' \, p_g(\mathcal{L}'),$$

the ROC curve is defined to be:

$$\mathrm{ROC} = \Sigma_g\left(\Sigma_q^{-1}(x)\right), \text{ where } \Sigma_q^{-1}(x) \text{ is the}$$

inverse of the quark's cumulative distribution. What is the ROC, as a function of quark fraction x?

(d) The area under the ROC curve (AUC) is also an interesting discrimination metric, often used by an (actual) machine in a gradient descent algorithm. What is the AUC for the ROC curve in part (c)?

(2) Consider the number of jets as defined by the procedure introduced in exercise (2) of lecture 1. 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 <u>discrete</u> 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, as a function of the parameter y_{cut}. Does the AUC for this observable ever correspond to better discrimination than that for τ_d from exercise (1)?