Re-Normalizing Flows

Taming Logs and Perturbation Theory in QCD

Rikab Gambhir

In collaboration with Radha Mastandrea

Based on: [**RG**, Mastandrea; 2XXX.XXXXX]

Perturbation Theory in QCD

Problem: Fixed-order perturbation theory in QCD generically suffers from non-physical issues that do not occur in real life!

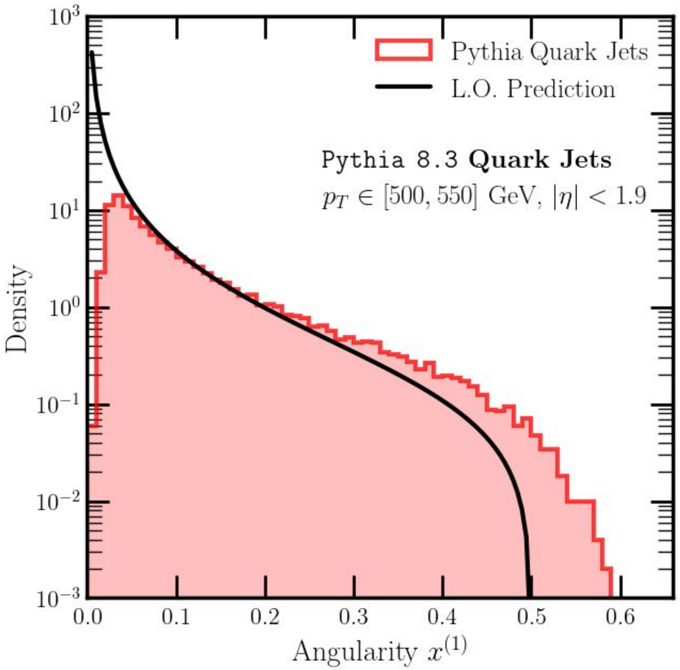
1. Non-finite
2. Non-smooth
3. Non-positive
4. Non-normalizable
5. Non-converging

Jet Angularity: $x^{(\beta)} = \sum_i z_i \left(\frac{\theta_i}{R}\right)^{\beta}$

$$p(x^{(1)}) = \frac{\alpha_s(xE_0)C_F}{2\pi}\int_0^R \frac{d\theta}{\theta}\int_0^1 dz\, P_{q\to qg}(z)\delta(x - z\frac{\theta}{R})$$

$$\sim \frac{\alpha_s(xE_0)C_F}{2\pi}\frac{R\log\left(\frac{R}{x}\right)}{x}$$

Fixed Order vs. Real*



*By "Real", I mean *PYTHIA* as a stand-in for real data

Everywhere I say "L.O." or "L.L", I am going to include running couplings. This will not change any point I make here

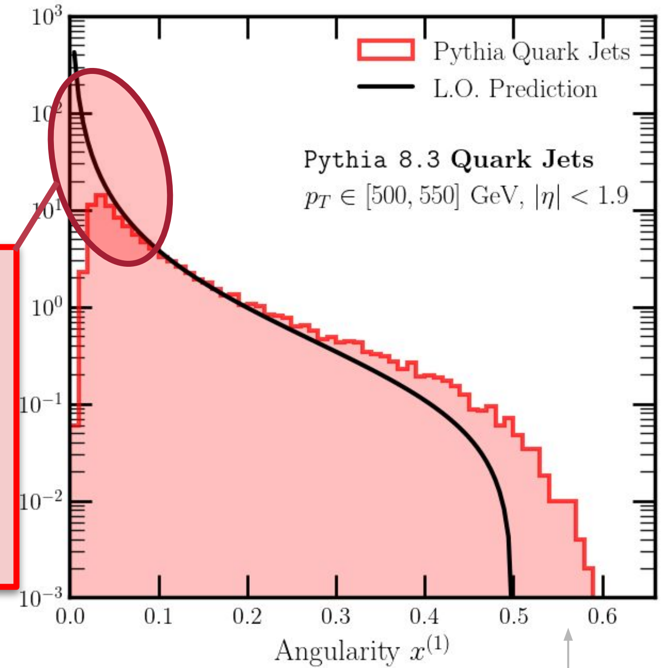
Perturbation Theory in QCD

Problem: Fixed-order perturbation theory in QCD generically suffers from non-physical issues that do not occur in real life!

1. Non-finite
2. Non-smooth
3. Non-positive
4. Non-normalizable
5. Non-converging

Logs blow up! This distribution is not normalized*, and convergence is spoiled when *x* is small!

Fixed Order vs. Real*



This difference is a genuine N.L.O effect, due to the definition of jet axis through the WTA algorithm.

Jet Angularity: $x^{(\beta)} = \sum_i z_i \left(\frac{\theta_i}{R}\right)^\beta$

$$p(x^{(1)}) = \frac{\alpha_s(xE_0)C_F}{2\pi}\int_0^R \frac{d\theta}{\theta}\int_0^1 dz\, P_{q\to qg}(z)\delta(x - z\frac{\theta}{R})$$

$$\sim \frac{\alpha_s(xE_0)C_F}{2\pi}\frac{R\log\left(\frac{R}{x}\right)}{x}$$

*Technically, these are *plus Distributions*, which are normalized. However, these are invisible to any real value of *x*, are highly non-smooth, and can be negative.

Perturbation Theory in QCD

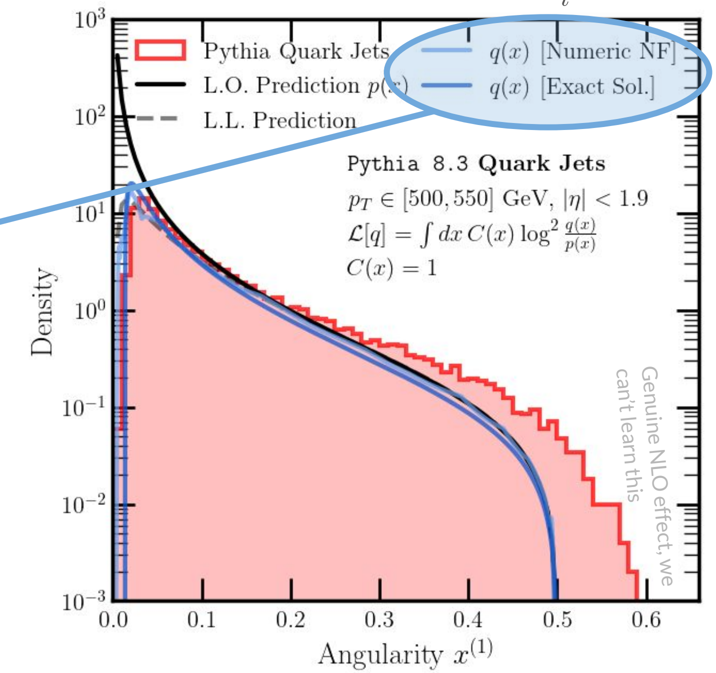
Jet Angularity: $x^{(\beta)} = \sum_i z_i \left(\frac{\theta_i}{R}\right)^\beta$

Problem: Fixed-order perturbation theory in QCD generically suffers from non-physical issues that do not occur in real life!

1. Non-finite → Finite!
2. Non-smooth → Smooth!
3. Non-positive → Positive!
4. Non-normalizable → Normalized!
5. Non-converging

Guaranteed by NF

I won't solve this one today, sorry!



Solution: *Force* **(1)-(4)** by matching the perturbation theory to a **Normalizing Flow (NF)**, which parametrizes the set of *all* finite, smooth, positive, normalizable functions, *while preserving the perturbative structure in α_s*!

Original, badly behaved prediction

$p(x) \longrightarrow q(x)$

New, well behaved prediction

Parametrized by NF

"Matching" Loss Functional $\mathcal{L}[q]$

The large logs have been tamed!

[Ellis, Hornig, Lee, Vermilion, Walsh; 1001.0014]
[Hornig, Lee, Ovanesyan; 0901.3780]

Even at NLO!

At NLO, can express using SCET by convolving:

$$f(a) \equiv \frac{1}{1-a/2}\left(\frac{7-13a/2}{4} - \frac{\pi^2}{12}\frac{3-5a+9a^2/4}{1-a} - \int_0^1 \mathrm{d}x \frac{1-x+x^2/2}{x}\ln[(1-x)^{1-a}+x^{1-a}]\right)$$

$$J_a^n(\tau_a^n;\mu) = \delta(\tau_a^n)\left\{1 + \frac{\alpha_s C_F}{\pi}\left[\frac{1-a/2}{2(1-a)}\ln^2\frac{\mu^2}{Q^2} + \frac{3}{4}\ln\frac{\mu^2}{Q^2} + f(a)\right]\right\} - \frac{\alpha_s C_F}{\pi}\left[\left(\frac{3}{4}\frac{1}{1-a/2} + \frac{2}{1-a}\ln\frac{\mu}{Q(\tau_a^n)^{1/(2-a)}}\right)\left(\frac{\theta(\tau_a^n)}{\tau_a^n}\right)\right]_+$$

$$S_a^{\mathrm{PT}}(\tau_a^s;\mu) = \delta(\tau_a^s)\left[1 - \frac{\alpha_s C_F}{\pi(1-a)}\left(\frac{1}{2}\ln^2\frac{\mu^2}{Q^2} - \frac{\pi^2}{12}\right)\right] + \frac{2\alpha_s C_F}{\pi(1-a)}\left[\frac{\theta(\tau_a^s)}{\tau_a^s}\ln\frac{\mu^2}{(Q\tau_a^s)^2}\right]_+$$

$$H(Q;\mu) = 1 - \frac{\alpha_s C_F}{2\pi}\left(8 - \frac{7\pi^2}{6} + \ln^2\frac{\mu^2}{Q^2} + 3\ln\frac{\mu^2}{Q^2}\right)$$

$$\frac{1}{\sigma_{\mathrm{tot}}}\frac{\mathrm{d}\sigma}{\mathrm{d}e} = H(Q;\mu)\int \mathrm{d}e_1\,\mathrm{d}e_s\, J_1(e_1;\mu)S(e_s;\mu)\delta(e-e_1-e_s)$$

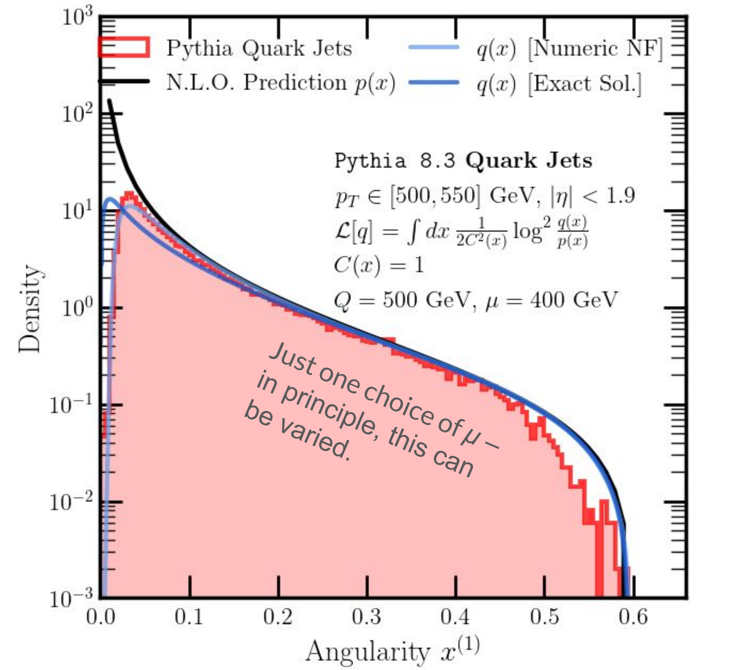
Large logs in *both* the angularity and in μ

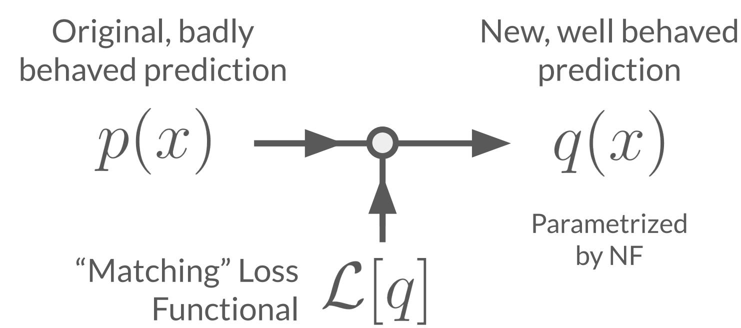
The large logs in angularity have been tamed!

Caveat: I cannot numerically promise matching to order α_s^2, only α_s^1. More details in the rest of the talk!

For the rest of this talk, I will stick with L.O for simplicity

Jet Angularity*: $x^{(\beta)} = \sum_i z_i \left(\frac{\theta_i}{R}\right)^\beta$





Ask me later how it may be possible to tame large logs in μ too!

*Technically a different jet angularity, but our point still stands

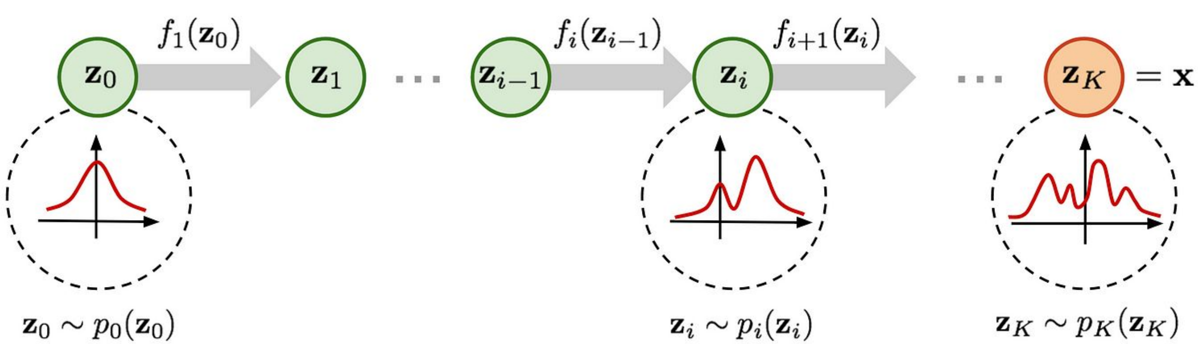
[Beate Sick, Hothorn, Dürr; 2004.00464]
[Ramasinghe, Fernando, Khan, Barnes; 2102.03509]
[Image from: Medium]

[Aside, if there's time] Normalizing Flows

A **Normalizing Flow** is a parameterized function *q(x)* that is *guaranteed* to be a valid probability distribution and can be easily sampled from.

This is accomplished by taking an *already known* probability distribution, $q_0(z)$, then parameterizing a transformation $x = f(z)$ that is invertible, has tractible Jacobians, and is easy to compose:

$$p_X(x) = p_Z(f^{-1}(x))|\det\frac{\partial f}{\partial z}|^{-1}_{z=f^{-1}(x)}$$



We choose to use *Bernstein-Polynomial Flows (BPFs)*, which we have found to be a nice, stable parameterization for low-dimensional flows.

Really, any method of learning distributions would work for our purposes! But NFs and in particular BPFs are easy.

Loss Matching

Given p, we want to find a q that still retains the salient physics of p – for example, matching the perturbative expansion of p, or matching the regions of p where the logs are not too large.

Encode this in a *Loss Functional*:

f is a choice. Here, we will pick *f* = *linear* or *log*.
This tells us in what space our "errorbars" are Gaussian in!

The dagger matters; *f* might be complex! See Backup Slides.

$$\mathcal{L}[q] = \int dx dy \left[f(q(x)) - f(p(x))\right]^\dagger \underbrace{\left[\frac{1}{2C^2(x,y)}\right]} \left[f(q(y)) - f(p(y))\right]$$

Gaussian Kernel: Tells us our tolerable "Gaussian Error" on matching! Choose this to regulate divergences.

$\frac{1}{2C^2(x,y)}$ can be anything, *especially* a differential operator, but for convenience we will often choose a diagonal and real kernel:

$$\mathcal{L}[q] = \int dx \, \frac{|f(q) - f(p)|^2}{2C^2(x)}$$

Why this loss? Most other losses can be encoded this way! Non-uniform sampling can also be built into C

Loss Matching - The C Term

Think of C as a Gaussian errorbar on *f(q)* – convenient, since our loss is an MSE!

The function C is a **physical choice**. Every choice of C will result in a unique* *q*. Each of these is valid, but if you want *q* to have any physical meaning, so should your choice of C!

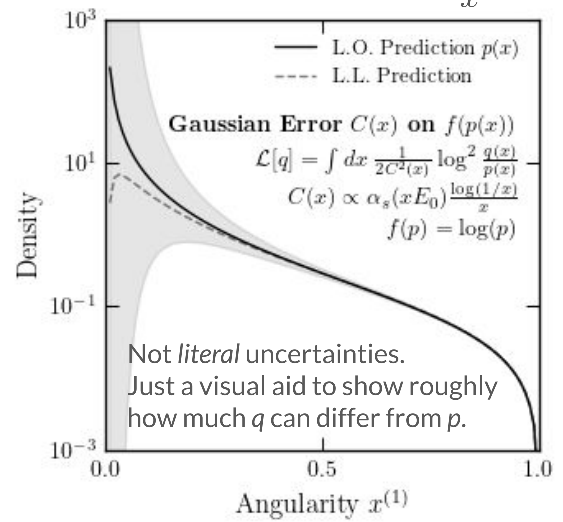
Need *(at least)* $[2C(x)]^{-1}$ to go to zero faster than *p(x)* blows up**.

e.g.
$$[2C(x)]^{-1} = \Theta(x-c)$$
$$[2C(x)]^{-1} = \frac{1}{p(x)}$$
$$[2C(x)]^{-1} = \frac{1}{\alpha_s + c\alpha_s^2 + \ldots}$$

Everything on this slide only applies to real, diagonal *C*. Later, *C* will be an abstract operator!
Note *C* is only defined up to an overall constant.

Example $f = \log$

$$C \propto \alpha_s \frac{\log(1/x)}{x}$$



For convenience, I've rescaled $x \to 2x$

$$\mathcal{L}[q] = \int dx\, \frac{|f(q) - f(p)|^2}{2C^2(x)}$$

*Uniqueness is only guaranteed where *C* has support
**Sufficient but not actually necessary

“Ok, but what C do I pick?”

Depends what specifically you want!

Few q's

I want a specific q: If you *already know what answer q* you want, just pick the magic C_q (defined shortly) – but of course, if you already know what q is, you are already done.

I want q containing all perturbative information in p: Choose C to be the **Taylor Expansion Operator** (defined shortly). *Extremely nontrivial to work with! But if it works, this is the best you can ever do!*

I just want to tame large logarithms: For f = log, choose C to preserve the exponential structure of the learned minimum.

I just want to encode some higher-order / all-orders information: Choose C proportional to α_s, and be careful that the Lagrange multipliers don't scale the wrong way!

I just want q to somewhat match p: Choose C to be large in the regions you trust p!

I just want q to regulate p: Sufficient to choose C to regulate divergences in p!

I don't care: Just initialize a random *NF*, and call it q!

Many q's

"Ok, but **what C do I pick?**"

Depends what specifically you want!

Few q's

I want a specific q: If you *already know what answer q* you want, just pick the magic C_q (defined shortly) – but of course, if you already know what q is, you are already done.

I want q containing all perturbative information in p: Choose C to be the **Taylor Expansion Operator** to order N:

$$[2C(x)]^{-1} = \exp_N[\alpha \frac{d}{d\alpha}|_{\alpha=0}] = \sum_{n=0}^{N} [\frac{1}{n!}\alpha^n \frac{d^n}{d^n\alpha}|_{\alpha=0}]$$

Very hard to minimize this differential operator! Function of α_s and x.

If this can be satisfied, this is *the* natural thing to choose!

Everything we know about $p(x)$

$$\exp[\alpha \frac{d}{d\alpha}]p(x) = p^0(x) + \alpha^1 q^1(x) + \ldots + \alpha^N p^N(x) + \mathcal{O}(\alpha^{N+1})$$

$$\exp[\alpha \frac{d}{d\alpha}]q(x) = q^0(x) + \alpha^1 q^1(x) + \ldots + \alpha^N q^N(x) + \mathcal{O}(\alpha^{N+1})$$

If these match: q has the *same* physics content as p to the given order, but is a valid distribution!

This C works generically for *any p* and *N*.

But we will see soon: Some easier choices of C work for particular p's at low Ns: in particular $C = 1$ works to order α_s^1

Many q's

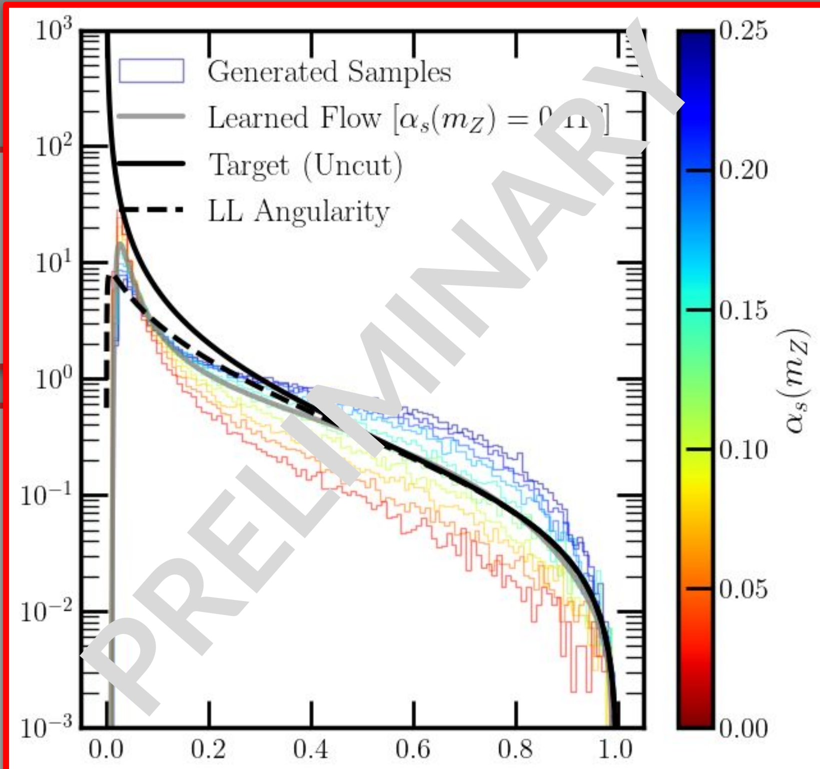
I don't care: Just initialize a random *NF*, and call it q!

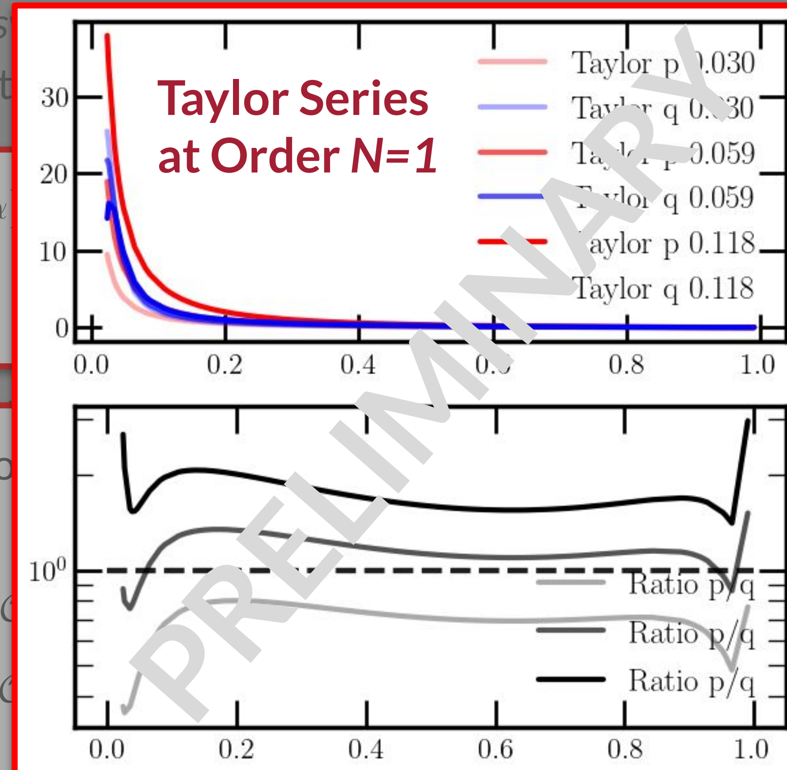
“Ok, but what C do I pick?”

$$[2C(x)]^{-1} = \exp_N[\alpha \frac{d}{d\alpha}|_{\alpha=0}] = \sum_{n=0}^{N}[\frac{1}{n!}\alpha^n \frac{d^n}{d^n\alpha}|_{\alpha=0}]$$

Few q's

Many q's





If these match: q has the same physics content as p to the

Learned distributions as a function of α look reasonable! But the Taylor expansion structure has not been preserved at order α_s^{1}.

This is a numerically challenging problem! Training dynamics are weird with losses of order 10^{30}.

Can this *ever* work?

Yes. If a perturbative answer exists, in principle, the NF should *eventually* find it!
For example, if we tell the NF to learn the *LL* resummation directly, it works exactly, and exists within the solution space! *This is a training dynamics problem, not a physics one!*

You, the user, give us:

If anybody at this workshop has ideas on doing this better, please tell us!

1. An expression $p^N(x|\alpha)$ representing a fixed-order expansion to order α^N.
   Or, samples *x* with weights auto-differentiable in α
2. A *promise* that $p^N(x|\alpha)$ is a fixed-order expansion of some "platonic" $p(x|\alpha)$ which is a non-pathological distribution.
   e.g. $p^N(x|\alpha)$ is an approximation to $p(x|\alpha)$ computed from the full QCD path integral

Ideally, we give you back:

A new distribution $q(x|\alpha)$ that is ***guaranteed*** to be normalized, smooth, positive, and finite, *and* **numerically** matches the perturbative expansion of $p^N(x|\alpha)$ to order *N*:

$$\exp[\alpha \frac{d}{d\alpha}]p(x) = p^0(x) + \alpha^1 q^1(x) + \ldots + \alpha^N p^N(x) + \mathcal{O}(\alpha^{N+1})$$

equal!

$$\exp[\alpha \frac{d}{d\alpha}]q(x) = q^0(x) + \alpha^1 q^1(x) + \ldots + \alpha^N q^N(x) + \mathcal{O}(\alpha^{N+1})$$

All of the information in *p(x)* is preserved!

Why is requirement [2], the promise, necessary? See backup slides or ask me later!

"Ok, but **what C do I pick?**"

Few q's

I want a specific q: If you *already know what answer q* you want, just pick the magic C_q (defined shortly) – but of course, if you already know what q is, you are already done.

I want q containing all perturbative information in p: Choose C to be the **Taylor Expansion Operator** (defined shortly). *Extremely nontrivial to work with! But if it works, this is the best you can ever do!*

Tactical Retreat to this region

Given that the perturbative expansion operator is *really hard*, let's just stick with real-valued C's for the rest of this talk. It turns out we can get at least part of the way there!

In particular, for the rest of this talk, $C(x)$ = some real function of x.

Many q's

*Almost everywhere.

Loss Matching – Exact Solutions

For f = log or linear and real-valued choices of C, we can directly solve for the optimal normalized, smooth*, finite, positive q by using Euler-Lagrange to minimize the MSE loss!

f = linear

$$q(x) = \mathrm{ReLU}\left(p(x) - \lambda C^2(x)\right)$$

f = log

$$q(x) = p(x)\exp[-W(\lambda p(x) C^2(x))]$$

Such that λ solves $\int dx\, q(x) = 1$

q(x) is a second-order correction to *p(x)*, especially if $C^2 \sim c(x)\alpha^2 + \ldots$

q(x) is an all-orders resummation of *p(x)*, especially if $C^2 \sim c_0(x) + c_1(x)\alpha + c_2(x)\alpha^2 + \ldots$

For general f: $q(x) = \mathrm{ReLU}\left[f^{-1}\left(f(p(x)) - \frac{\lambda C^2(x)}{f'(p(x))}\right)\right]$

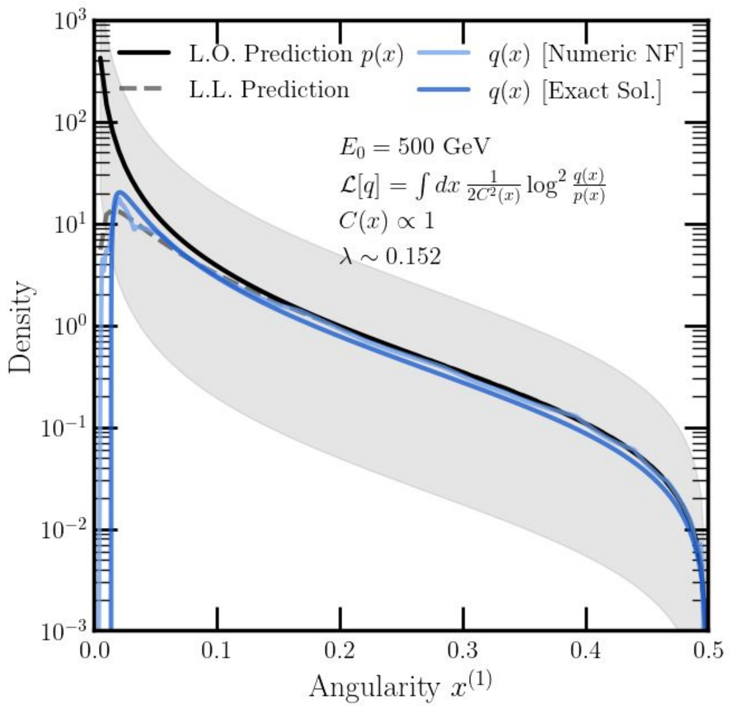
If we have exact solutions, why bother with *NF*s? We might not have closed forms for q, we might be working in many dimensions, and solving for λ is numerically hard.

Taming Large Logarithms – Example

Let's pick *f* = log, *C(x)* = 1 for the L.O. jet angularity!

$$
\begin{aligned}
q(x) &= p(x)\exp[-W(\lambda p(x) C^2(x))] \\
&= \left[\frac{\alpha_s C_F}{\pi}\frac{\log(1/x)}{x}\right]\exp\left[-W(\lambda\left[\frac{\alpha_s C_F}{\pi}\frac{\log(1/x)}{x}\right])\right] \\
&\approx \left[\frac{\alpha_s C_F}{\pi}\frac{\log(1/x)}{x}\right]\exp\left[-\frac{\lambda\alpha_s C_F}{\pi}\frac{\log(1/x)}{x} + \mathcal{O}(\alpha_s^2)\right] \\
&= p(x) + \mathcal{O}(\alpha_s^2)
\end{aligned}
$$

"Sudakov"-Like



Perturbative structure preserved at α!
Exponential suppression of the divergence!
Sudakov-like!

Due to running

Sudakov

$$
q^{LL}(x) = \left[\frac{\alpha_s C_F}{\pi}\frac{\log(1/x)}{x} + \frac{\beta_\alpha(\alpha)}{2\pi x}\log^2(1/x)\right]\exp\left[-\frac{\alpha_s C_F}{2\pi}\log^2(1/x) + \mathcal{O}(\alpha_s^2)\right]
$$

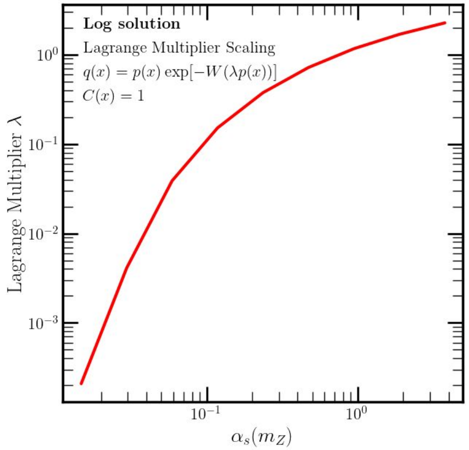
For comparison, the full L.L.

C = 1 is OK for *f* = log, but not linear. Ask me why later!

Taming Large Logarithms – Details

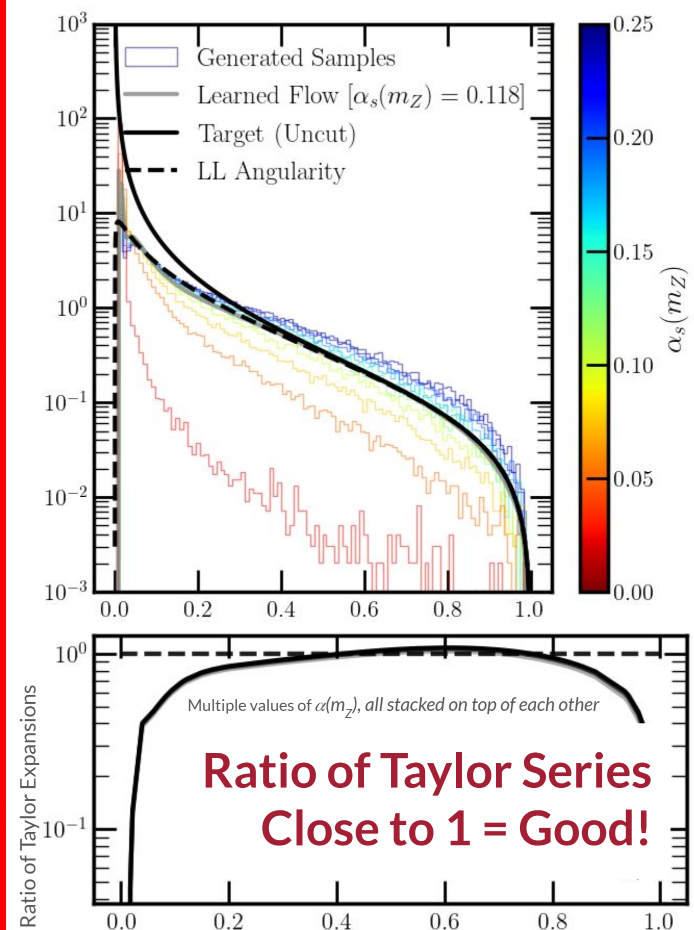
A catch! If the Lagrange Multiplier λ scales as α^{-1}, the perturbative structure is doomed! Let's check.

$$
\begin{aligned}
q(x) &= p(x)\exp[-W(\lambda p(x) C^2(x))] \\
&= \left[\frac{\alpha_s C_F}{\pi}\frac{\log(1/x)}{x}\right]\exp\left[-W(\lambda\left[\frac{\alpha_s C_F}{\pi}\frac{\log(1/x)}{x}\right])\right] \\
&\approx \left[\frac{\alpha_s C_F}{\pi}\frac{\log(1/x)}{x}\right]\exp\left[-\frac{\lambda\alpha_s C_F}{\pi}\frac{\log(1/x)}{x} + \mathcal{O}(\alpha_s^2)\right] \\
&= p(x) + \mathcal{O}(\alpha_s^2)
\end{aligned}
$$



Numerically, we are *ok*!
We can say *numerically* that *C(x) = 1* preserves perturbative structure at α^1!

If λ itself scales as α^1, we could go even further and say *C(x) = 1* preserves α^2 for some *p*, but this numerically does not seem to be the case, it seems to be a nontrivial power law



For comparison, the full L.L.

C = 1 is OK for *f* = log, but not linear. Ask me why later!

[Aside, if we have time] **Magic C Functions**

Similar conclusions hold for *f* = linear

For any valid *q*, there exists a choice of *C*, (call it the **magic C_q**) such that *p* corrects to *q*. For *f* = log:

$$\frac{1}{2C_q^2(x)} \propto \frac{q(x)}{2\log\frac{q(x)}{p(x)}} \longrightarrow \begin{aligned}\mathcal{L}[q] &= \int dx\, q(x)\log\frac{q(x)}{p(x)} \\ &= D_{KL}(q||p)\end{aligned}$$

The existence of magic *C*'s makes it clear that there is **no free lunch**: Unless you have a good reason to pick a particular C, there is infinite ambiguity in the final q and our method provides no genuinely new information, since any *q* is accessible! Can't avoid physics!

If you choose the magic C_q, the result of the loss functional minimization will be *q*. The *value* of the loss will be the KL divergence between *q* and *p*!

[Aside, if we have time] Magic C Functions

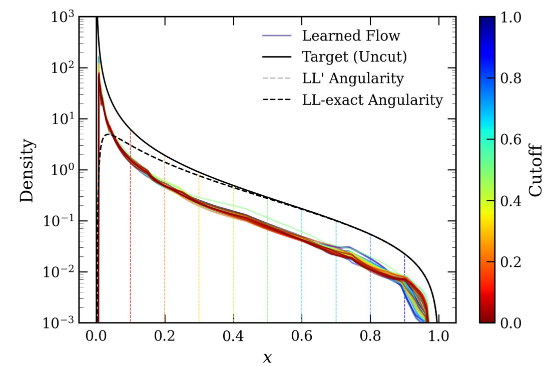
For $q(x) = q^{LL}(x)$ with *no running*:

$$\mathcal{L}[q] = \int dx\, q(x) \log \frac{q(x)}{p(x)}$$
$$= D_{KL}(q||p)$$
$$= \mathbb{E}_{x\sim q} \log^2(x)$$

Logarithmic moments arise naturally using the magic *C*'s! See first use in related work by [Assi, Höche, Lee, Thaler; 24XX.XXXX]

Generically, for any *q* we get the moment of some distinguished observable

What is the *best* possible *q*?
Equal to minimizing over all choices of C.
Not well-defined, but NF provides **inductive bias** and picks one!!



Optimal solution:
q = p X [infinite constant]
An NF can't learn this, but it has to learn *something*: *p ~ q*

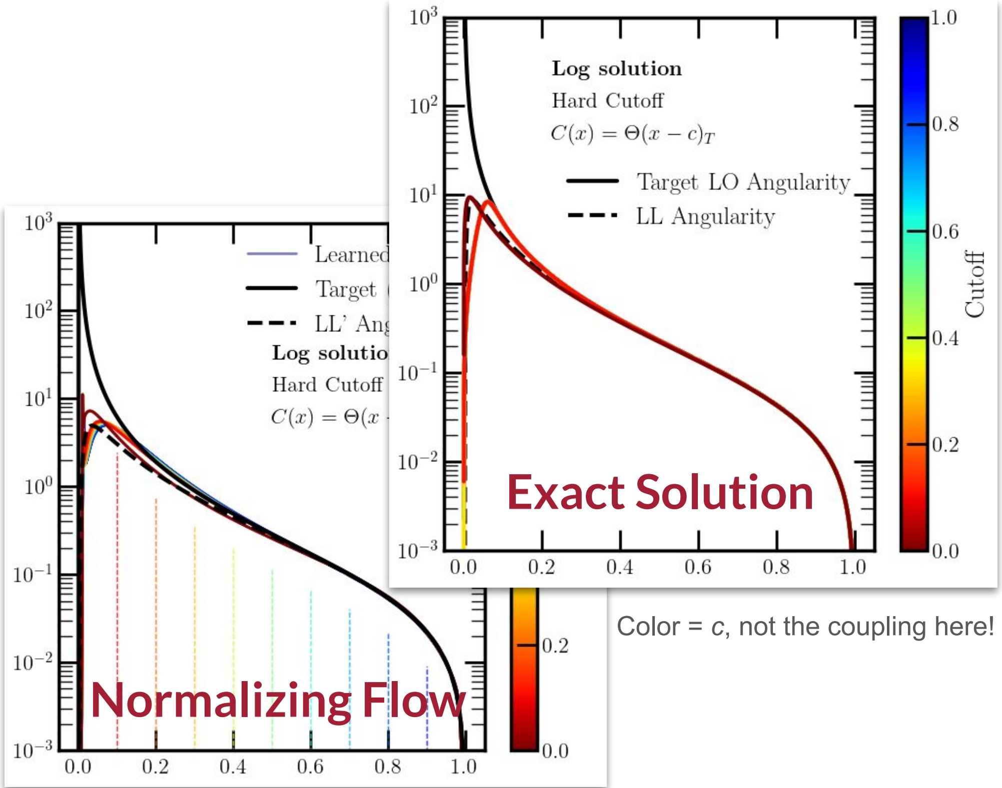
Plot Dump: More Choices of C

C(x) = 1 is *not* the only possible choice! By a similar argument, other choices of *c* also work, and result in slightly different predictions at order α^2.

$$[1/2C(x)]^{-1} = \Theta(x - c)$$

or some softer cutoff…

"I want *q* to match *p* to the right of some cutoff *c*, where I trust my perturbation theory more because logs are smaller"



Color = *c*, not the coupling here!

Choosing a soft cutoff *c* to be related somehow to the renormalization scale μ might be useful! Ask me about this later

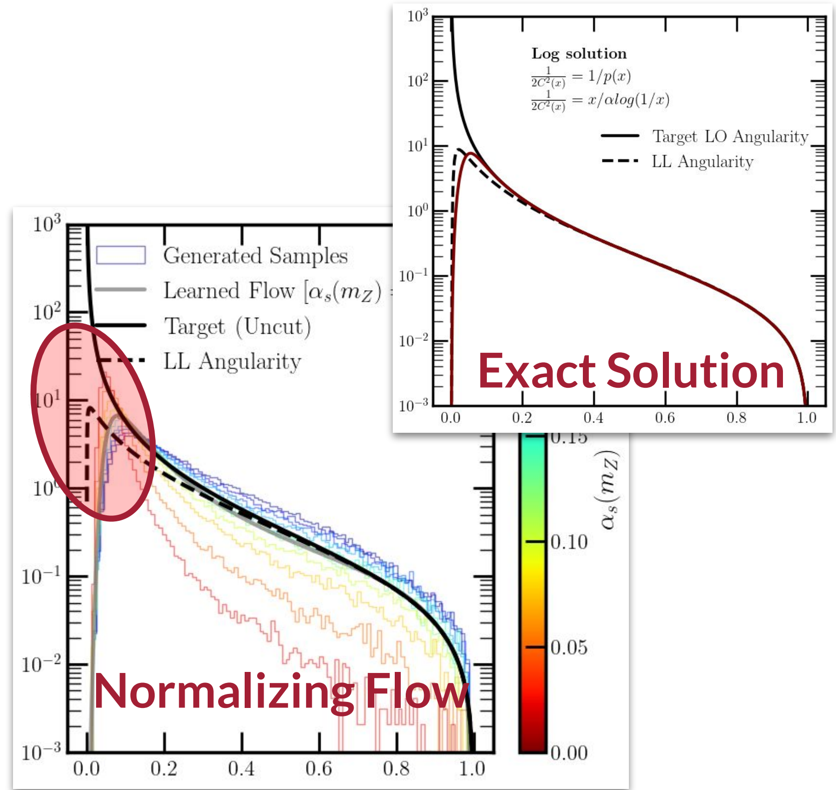
Plot Dump: More Choices of C

C(x) = 1 is *not* the only possible choice! By a similar argument, other choices of *c* also work, and result in slightly different predictions at order α^2.

$$[1/2C(x)]^{-1} = 1/p(x)$$
$$\propto x/\alpha_s \log(1/x)$$

"I want *q* to match *p* in the region where the order-1 coefficients in my perturbation theory are smaller"

Slightly different than *C = 1!*



CLAIMER

What we are claiming

- **Claiming:** Given a distribution *p*, pathological or otherwise, we can find a regulated distribution *q* that is normalized, smooth, positive, and normalized that is perturbatively close to *p*.
- **Claiming:** This is a universal approximator: *any* valid *q* can be reached from *any** invalid *p* with the correct choice of *C*.
- **Claiming:** For *particular* *C*'s, we can get meaningful *q*'s that look like theory calculations and tame large logs!

DISCLAIMER

What we are *not* claiming

- **Not Claiming:** The perturbative structure is *easy* to maintain numerically or that numeric artifacts are manageable.
  - But if you restrict to special choices of *C*, like $C(x) = 1$, you can maintain perturbative structure up to some finite order!
- **Not Claiming:** Varying *C*'s can give an "uncertainty envelope" on the space of *q*'s, *without* a prior on the space of *C*'s.
  - But if you *do* have a prior on a family of *C*'s, then this is fair game!
- **Not Claiming:** Any random choice of *C* will work.
  - There do exist *simple* *C*'s that do work, but you still have to know to choose them and know that they will not ruin perturbative structure!

Key Point: This is *not* a shortcut to doing QCD, just a new way to parameterize things to guarantee nice properties. There is no free lunch, you still have to do physics!

What we are still thinking about …

- The Taylor expansion operator $[2C(x)]^{-1} = \exp_N[\alpha \frac{d}{d\alpha}|_{\alpha=0}]$ is extremely hard to train. However, we know the solution should be possible and live within the NF solution space – is there a variant of this with better training dynamics?
- We only explored one dimensional distributions in this talk. However, in principle, nothing prevents us from considering multidimensional distributions, and NFs might be especially useful for this!
    - The "platonic ideal" of this is to use the full QCD phase space rather than just individual observables!
- We did not discuss using the loss to enforce regulator or renormalization scale independence numerically

Email me questions and comments at rikab@mit.edu
Or the ML4Jets Slack!

In principle possible!

$$p(x) = p^0(x) + \alpha^1 q^1(x) + \ldots + \alpha^N p^N(x) + \mathcal{O}(\alpha^{N+1})$$

$$q(x) = q^0(x) + \alpha^1 q^1(x) + \ldots + \alpha^N q^N(x) + \mathcal{O}(\alpha^{N+1})$$

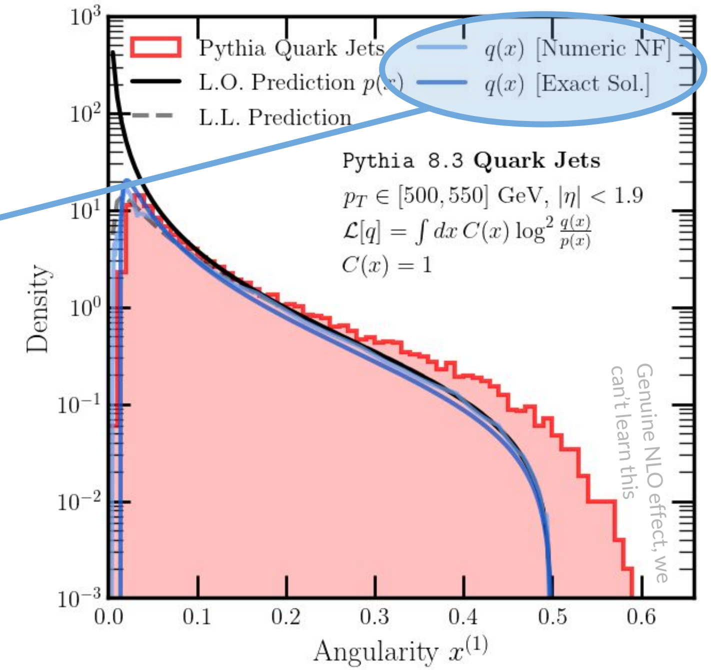
Numerically feasible!

Conclusion

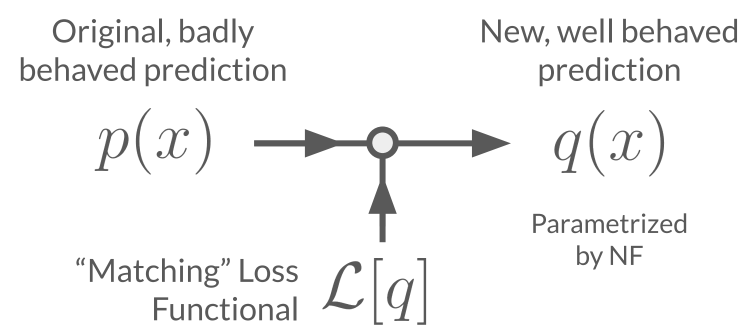
Problem: Fixed-order perturbation theory in QCD generically suffers from non-physical issues that do not occur in real life!

1. Non-finite → **Finite**!
2. Non-smooth → **Smooth**!
3. Non-positive → **Positive**!
4. Non-normalizable → **Normalized**!

Guaranteed by NF



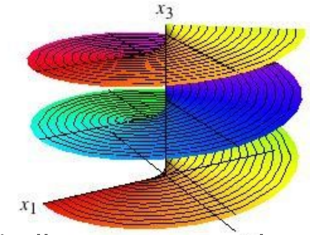
Solution: *Force* (1)-(4) by matching the perturbation theory to a **Normalizing Flow (NF)**, with the matching conditions encoded in the C function, with the hope of the preserving perturbative expansion.



The large logs have been tamed!

Backup

[Joshua 1:9]



Biblically accurate complex plane

“Logs? But what if *p* is negative?”

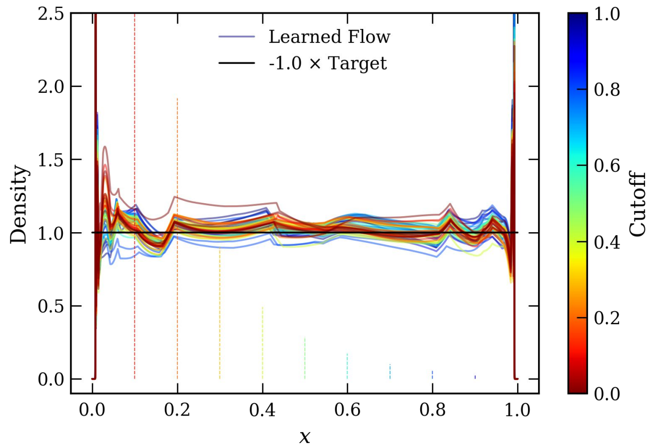
Be not afraid!

From $\log(p) = \log(|p|) + i\arg(p)$

$$\mathcal{L}[q] = \int dx \, \frac{|\log(q) - \log(p)|^2}{2C^2(x)} = \int dx \, \frac{\log^2\left|\frac{q}{p}\right| + \frac{\log^2(q)}{\pi^2}\Theta(-p(x))}{2C^2(x)}$$

“Make *q* as close to zero possible to match the negative *p* without going negative”

Explicit example for *p(x) = -1, C(x) = 1*

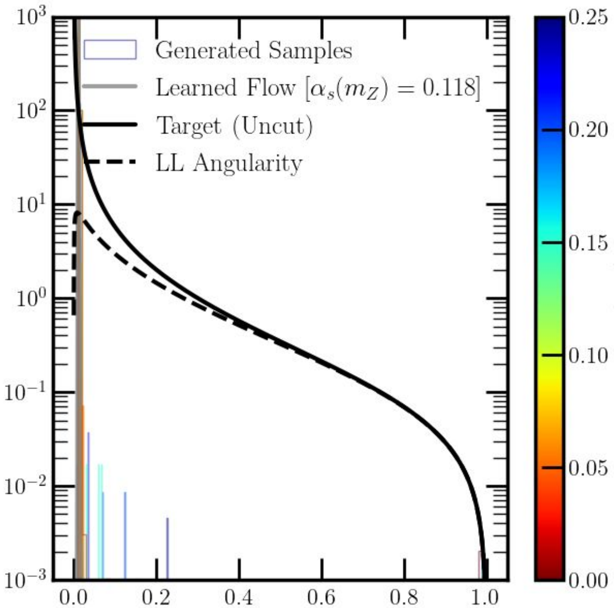


The real caveat is that *p(x)* cannot be *0*, at least on any extended region of phase space

What goes wrong with the Taylor Expansion?

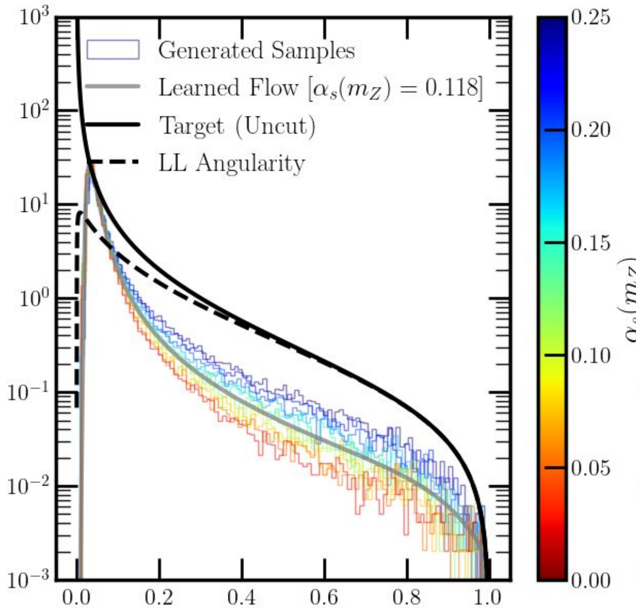
All of these networks are “pretrained” with the *C = 1* loss.

Sometimes, it learns a delta function!



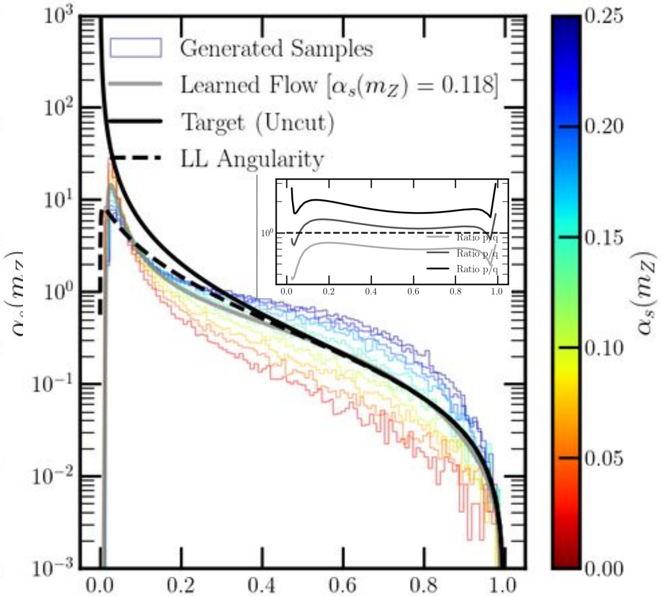
L ~ [log q/p] [$\exp_N$(α d/dα)][log q/p]

Sometimes, it learns a terrible match!!



L ~ [log $p_0/q_0]^2$ + [log $p_1/q_1]^2$

Sometimes, it looks ok, but the Taylor expansion is very off



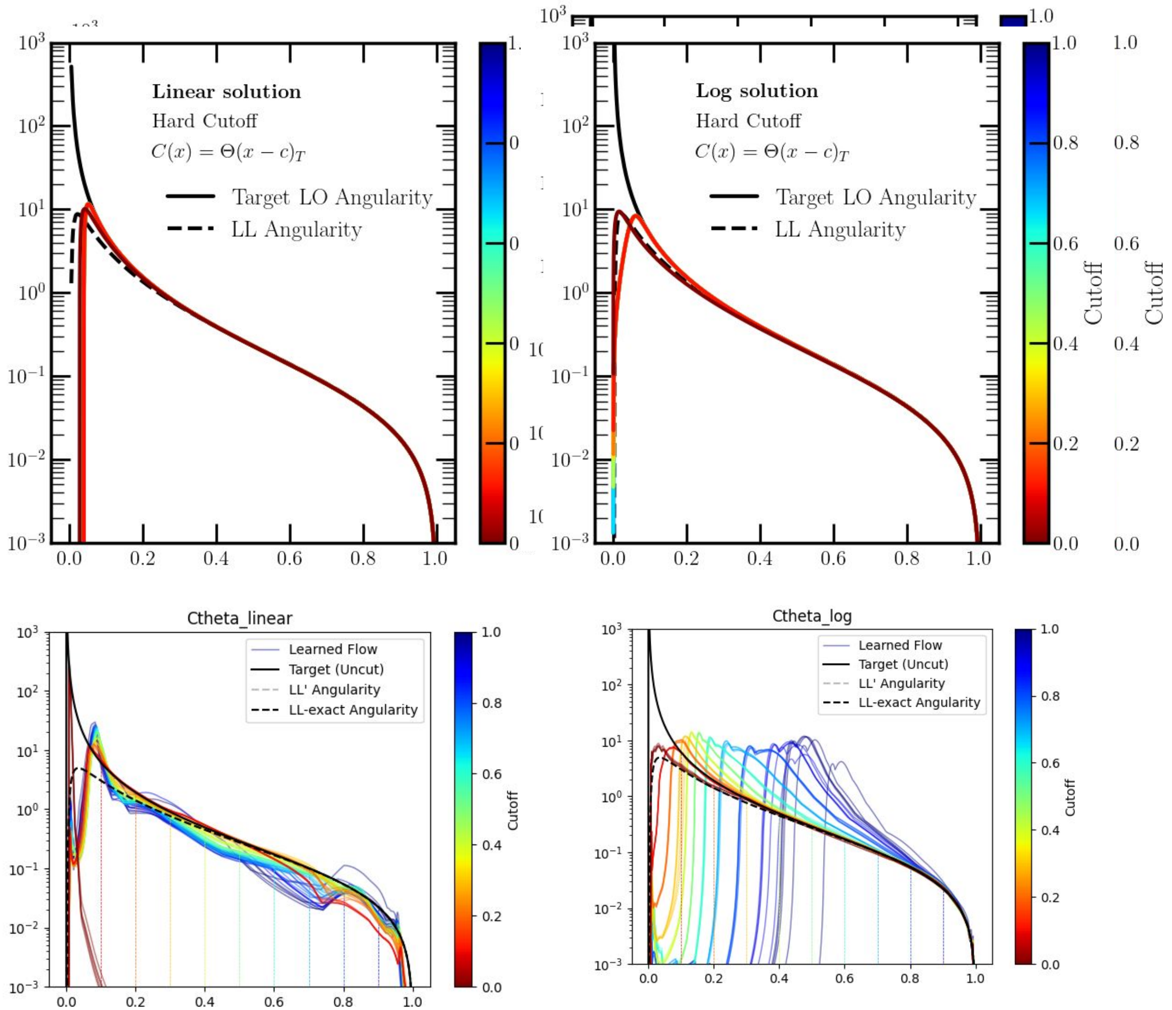
L ~ [$\exp_N$(α d/dα)][log q/p]^2

Plots: Cutoff $C(x)$ = *Hard Cutoff* *with sigmoid regularization*

Linear | Log

Exact

Numeric / NF

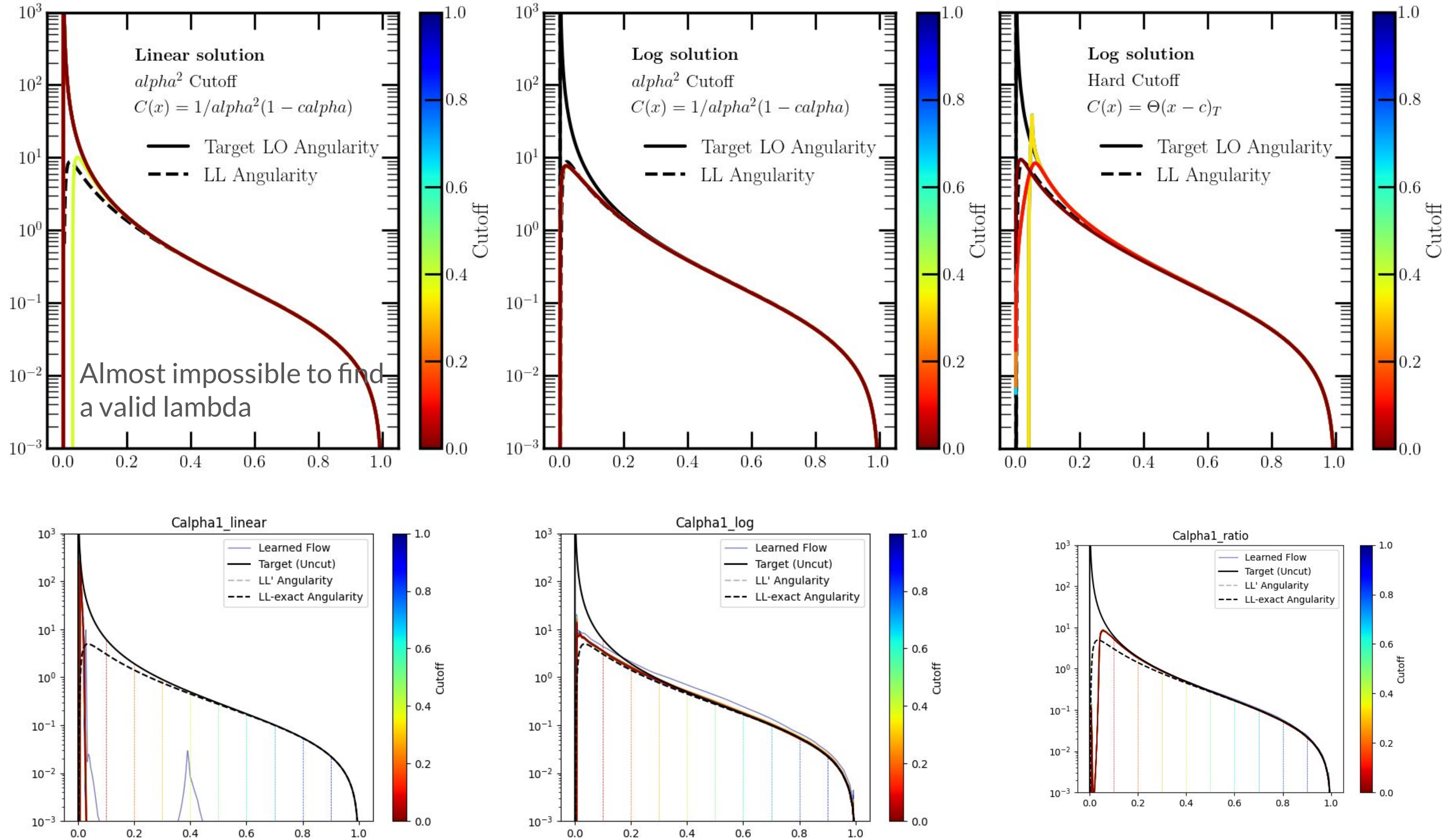


Plots: Cutoff *C(x) = 1/(alpha^2 + c alpha^3)*

Linear | Log | Ratio-MSE

Exact

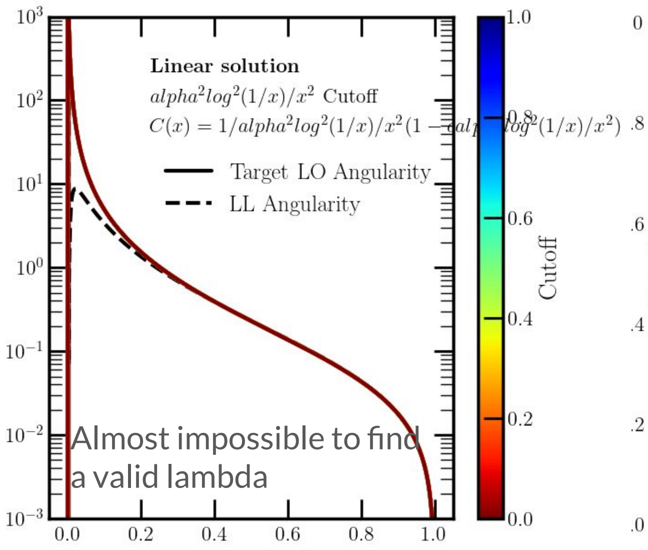
Numeric / NF

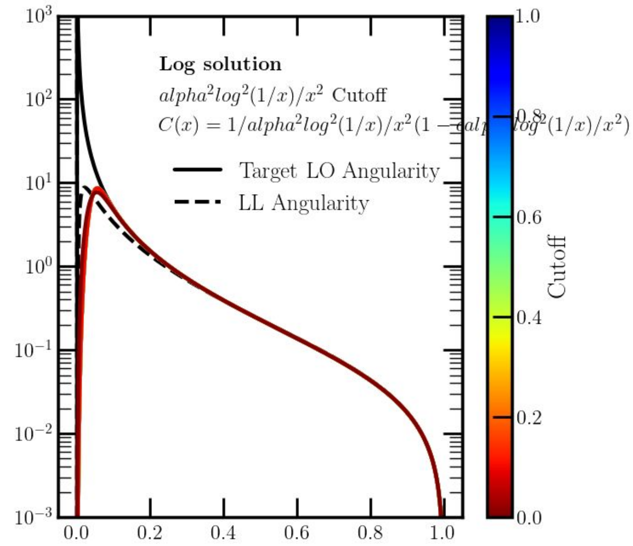


Plots: Cutoff *C(x) = 1/(alpha^2 + c alpha^3)logs*

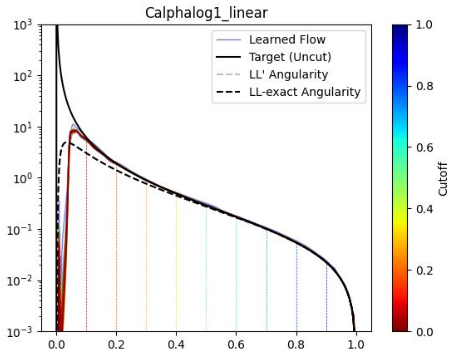
Linear | Log

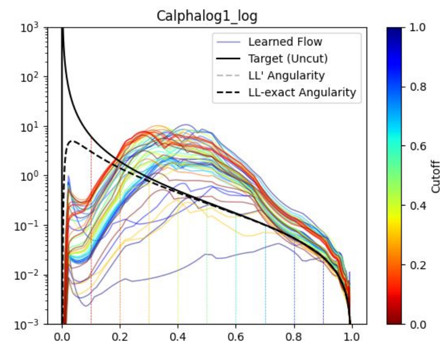
Exact





Numeric / NF





Network specifications

Networks are BPFs (Bernstein-Polynomial Flows) implemented with probabilists/zuko

Networks have 5 BPF blocks. Flow components have 2 layers each with 32 hidden features. In total, there are 8405 trainable parameters.

Networks are trained for 2000 epochs with a learning rate of 1e-3 and a batch size of (512 x choices) * (32 α choices). Evaluation is done at the final epoch.