Towards Machine Learning Analytics for Jet Substructure

arXiv:2007.04319

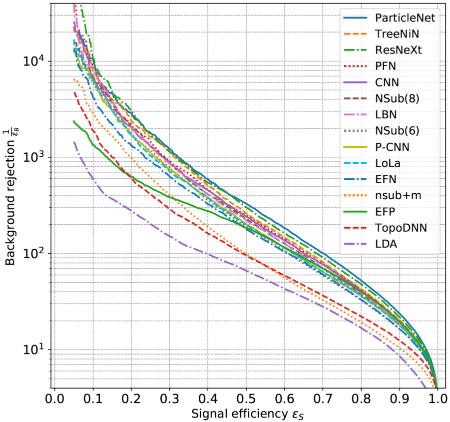
Giovanni Stagnitto
with G. Kasieczka, S. Marzani and G. Soyez

LPTHE & Université Paris Diderot
INFN & Università degli Studi di Milano

BOOST 2020 - July 21

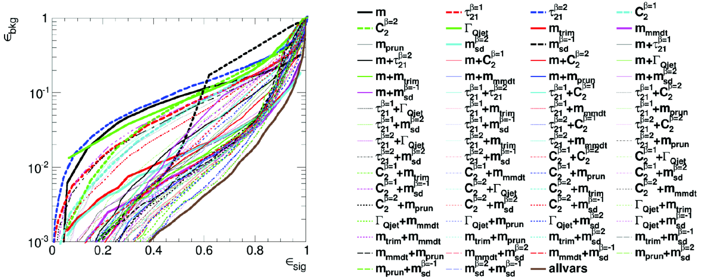
Introduction and motivations

BOOST 2019 (top tagging):



(refer to the paper for references)

BOOST 2013 (W tagging):



Deeper understanding of jet substructure algorithms reached only with QCD analytical studies

ML for jet substructure:
first-principle understanding possible?

Quark/gluon tagging with N-subjettiness

General approach put forward in [Larkoski, Metodiev (2019)], where $\{\tau_1 \dots \tau_n\}$ are adopted to resolve n emissions in the jet.

- Calculate the probability distribution for background (gluons) and signal (quark) jets:

$$p_i(\tau_1, \dots, \tau_n) = \frac{1}{\sigma_i} \frac{d\sigma_i}{d\tau_1 \cdots d\tau_n}, \quad i = q, g$$

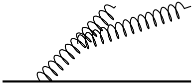
  (done at LL accuracy for $n = 1, 2, 3$).

- Find likelihood ratio, best single-variable discriminant, as:

$$\mathcal{L}(\tau_1, \dots, \tau_n) = \frac{p_{\mathcal{B}}}{p_{\mathcal{S}}} = \frac{p_g(\tau_1, \dots, \tau_n)}{p_q(\tau_1, \dots, \tau_n)};$$

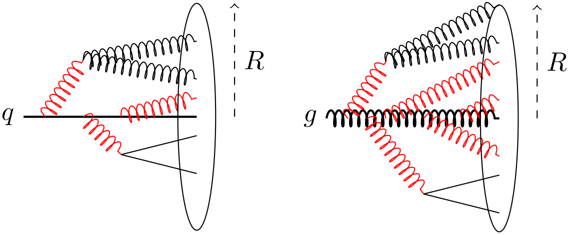
- Eventually calculate cumulative distribution for a cut on $\mathcal{L}$, ROC and AUC.

Even at LL accuracy, p_i has a **complicated structure** due to subsequent gluon splittings.



Primary N-subjettiness

Introduce a variant of τ_n that is **sensitive, at LL accuracy, only to primary emissions**.



Set of (z_i, Δ_i), for $i = 1, \ldots, m$, that we order such that $z_1 \Delta_1^\beta \geq z_2 \Delta_2^\beta \geq \cdots \geq z_m \Delta_m^\beta$.

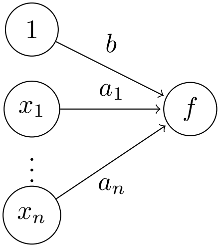
Definition: $$\mathcal{T}_N = \sum_{i=N}^{m} z_i \left(\frac{\Delta_i}{R}\right)^\beta$$

At LL: $$p_i(\mathcal{T}_1, \ldots, \mathcal{T}_n) = \left(\frac{\alpha_s}{\pi\beta}\right)^n (2C_i)^n \prod_{j=1}^{n} \left(\frac{\log(1/\mathcal{T}_j)}{\mathcal{T}_j}\right) \exp\left[-\frac{\alpha_s}{\pi\beta} C_i \log^2 \mathcal{T}_n\right]$$ (here fixed α_s)

- p_i is the *same* for q and g jets except for the colour factor, C_F or C_A.
- At LL, **a cut on the likelihood ratio** $\mathcal{L} = p_g/p_q$ **is equivalent to a cut on** $\mathcal{T}_n$.
- Possible to find analytic expressions for cumulative distribution and AUC for any n.

Perceptron analytics

Use primary N-subjettiness as input to the simplest possible NN, just one neuron: $y = f(\vec{a} \cdot \vec{x} + b)$



The best choice of the weights $\vec{a}$ and b is learnt by the perceptron by minimising the cost function (with f sigmoid and C cross-entropy):

$$\widetilde{C}(\vec{a}, b) = \frac{1}{2} \int d\vec{x} \left[p_q(\vec{x})\, C(f(\vec{x} \cdot \vec{a} + b), 0) + p_g(\vec{x})\, C(f(\vec{x} \cdot \vec{a} + b), 1) \right)$$

Do the weights at the mininum of the cost function correspond to optimal discriminant, which is a cut on x_n i.e. $a_1 = \cdots = a_{n-1} = 0$?

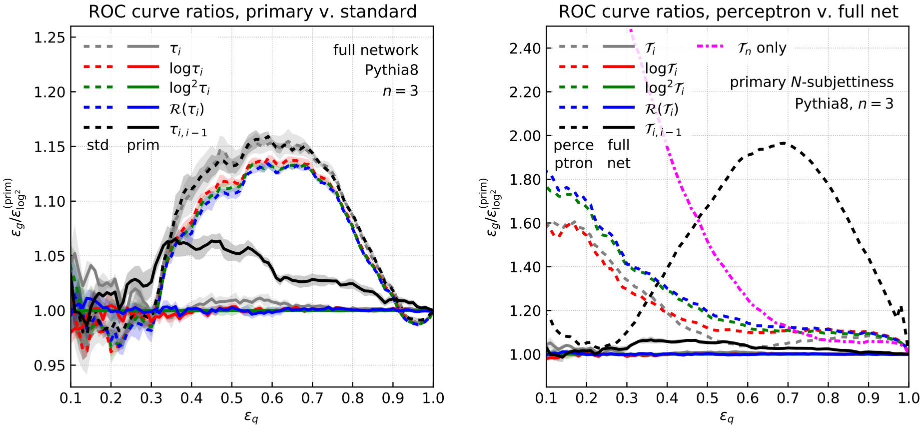
It depends on the functional form of the input variables:

- ratios $\mathcal{T}_{n,n-1} = \mathcal{T}_n / \mathcal{T}_{n-1}$: ✗;
- logarithmic inputs ($x_i = \log^2 \mathcal{T}_i$ or $x_i = \log \mathcal{T}_i$): ✓;
- linear inputs ($x_i = \mathcal{T}_i$): ✗ (altough within the reach of the network).

Analytical findings validated with an actual implementation of the perceptron.

Monte-Carlo studies

Extend our setup: fully-fledged NN trained with Monte-Carlo (Pythia) pseudo-data.



- **Primary N-subjettiness better** for the most of the range in quark efficiency.
- For $\varepsilon_q \gtrsim 0.6$, perceptron only 10% worse than the full NN, and similar to a simple cut on $\mathcal{T}_n$.
- Full NN achieves a similar degree of performance regardless of the type of inputs, but **more training or fine-tuning required in case of linear inputs or N-subjettiness ratios**.

Take-home messages

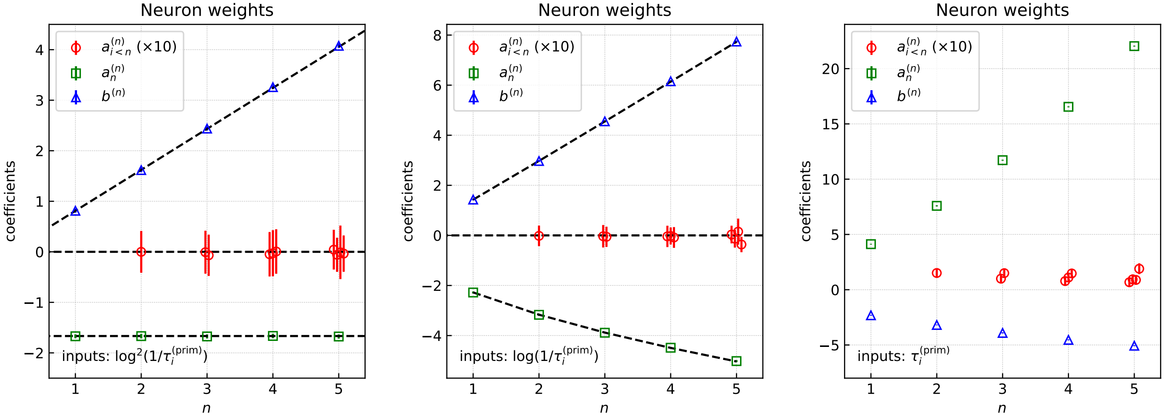
Positive role of first-principle understanding of the underlying phenomena in classification problems that employs ML techniques.

- ***Primary*** **N-subjettiness** $\mathcal{T}_N$, more amenable to all-order QCD analysis, while maintaining, if not exceeding, the discriminating power.
  $\rightarrow$ Going beyond LL? Primary N-subjettiness in vector boson or top tagging?
- **At LL, the optimal discriminant is just a cut on** $\mathcal{T}_n$. A perceptron is able to find the correct minimum if $\log^2 \mathcal{T}_i$ or $\log \mathcal{T}_i$ are passed, but fails to do so with $\mathcal{T}_i$.
  $\rightarrow$ Analytical study of more complex networks?
- **Qualitative agreeement with full NN trained with Monte-Carlo pseudo-data**. Full NN achieves a similar degree of performance regardless of the type of inputs, but more training or fine-tuning required in case of linear inputs or N-subjettiness ratios.

Backup slides

Perceptron numerics

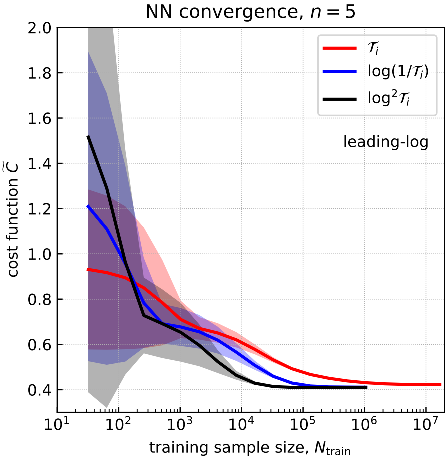
Validate analytic finding with an actual implementation of the perceptron (with n inputs $\mathcal{T}_i$), trained with a sample of pseudo-data generated according to the QCD LL distribution.

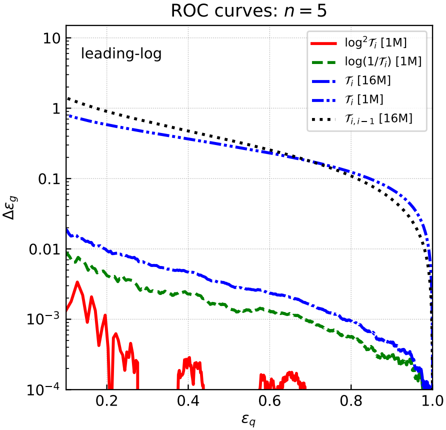


Visible effect also on the NN convergence (slower with linear inputs)

Perceptron numerics

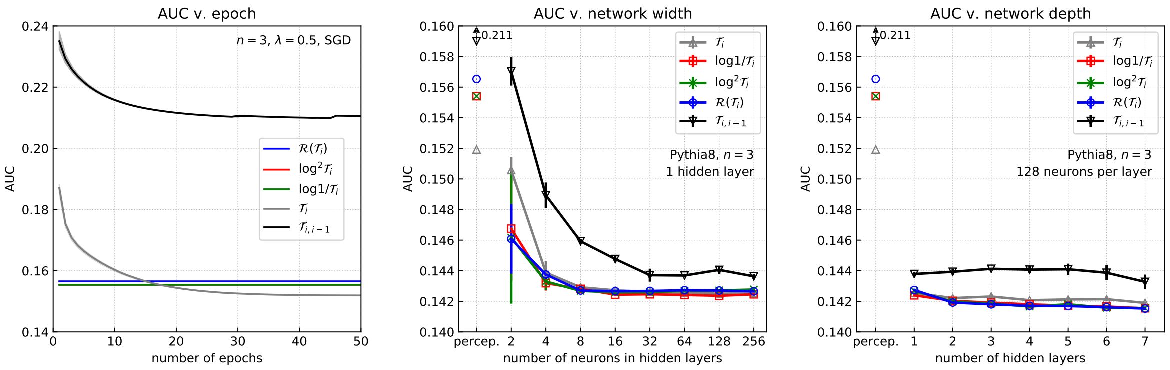
Do the theoretical issues for the linear inputs have some visible effect on the NN performance?





$$\Delta\epsilon_g = \frac{\text{ROC}_{NN}}{\text{ROC}_{\text{lik.}}} - 1$$

Monte-Carlo studies



Monte-Carlo studies