ML4Jets Rutgers 2022

Conditional generative networks for pure quark and gluon jets

Based on arXiv:2211.xxxxx

Ayodele Ore

ayodeleo@student.unimelb.edu.au

The University of Melbourne

Quarks and gluons from data

- Distinguishing quarks and gluons is a key task at LHC:
  - New physics discovery
  - α_S measurements
  - PDF determination
- ML models trained on MC parton labels are powerful but
  - Subject to QCD or detector mismodelling.
  - Q/G labels are ambiguous at the detector level.
- Data-driven methods (trained on Q/G mixtures) can evade these issues.

Jet topics Metodiev et al. [1802.00008]

1. Given two mixtures

$$p_{M_1}(x) = f_1\, p_Q(x) + (1 - f_1)\, p_G(x)$$
$$p_{M_2}(x) = f_2\, p_Q(x) + (1 - f_2)\, p_G(x)$$

2. Determine *reducibilities*

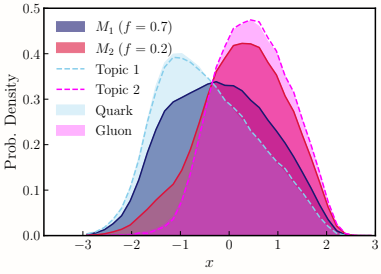
$$\kappa_{ij} = \min_x \frac{p_{M_i}(x)}{p_{M_j}(x)}$$

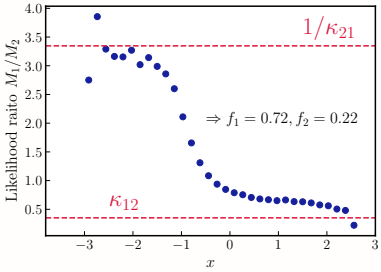
3. Recover fractions and distributions

$$f_1 = (1 - \kappa_{12})/(1 - \kappa_{12}\kappa_{21}), \quad f_2 = \kappa_{21} f_1$$

$$p_Q(y) = \frac{p_{M_1}(y) - \kappa_{12}\, p_{M_2}(y)}{1 - \kappa_{12}}, \quad p_G(y) = \frac{p_{M_2}(y) - \kappa_{21}\, p_{M_1}(y)}{1 - \kappa_{21}}$$

- Accurate and interpretable.
- Binning makes likelihood evaluation and sampling expensive in high dimension.





Deep-generative approach

How could a generative model for p_Q and p_G be trained?

- Normalizing flow with target likelihood
  - Train flows for M_1 and M_2
  - Use κ_{ij} and learned p_{M_k} to construct target
  - Train new flows with reverse-KL
- Event subtraction GAN
  - Re-purpose GAN setup from Butter et al. [1912.08824]
- **Conditional model**
  - Train single model on M_i with f_i as condition
  - Generate p_Q and p_G using conditions $f = 1, 0$

$$p_Q(x) = \frac{p_{M_1}(x) - \kappa_{12}\, p_{M_2}(x)}{1 - \kappa_{12}}$$

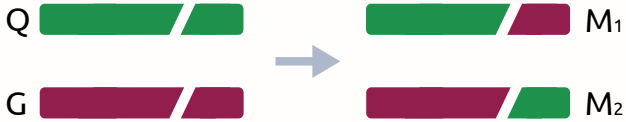
$$p_G(x) = \frac{p_{M_2}(x) - \kappa_{21}\, p_{M_1}(x)}{1 - \kappa_{21}}$$

$$p_{M_i}(x) = p(x|f_i)$$

$$p_Q(x) \equiv p(x|1) \quad p_G(x) \equiv p(x|0)$$

Datasets

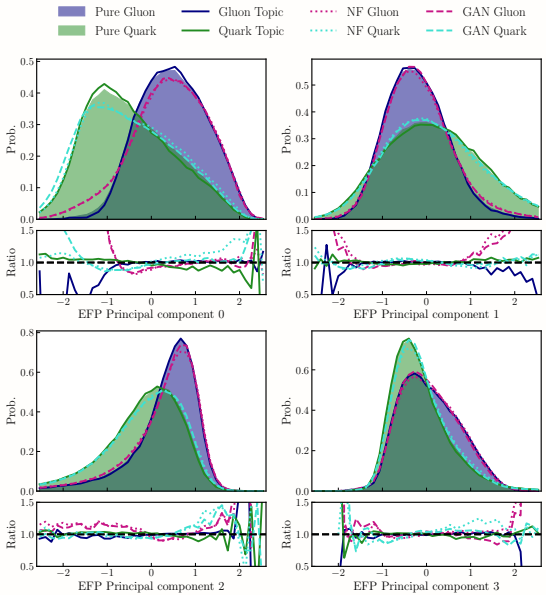
- PYTHIA quark/gluon dataset from `EnergyFlow`
  - 2M jets total
  - $R = 0.4$ anti-k_T
  - $p_T \in [500, 550]$ GeV
- Quark/gluon mixtures are constructed by splicing full dataset:



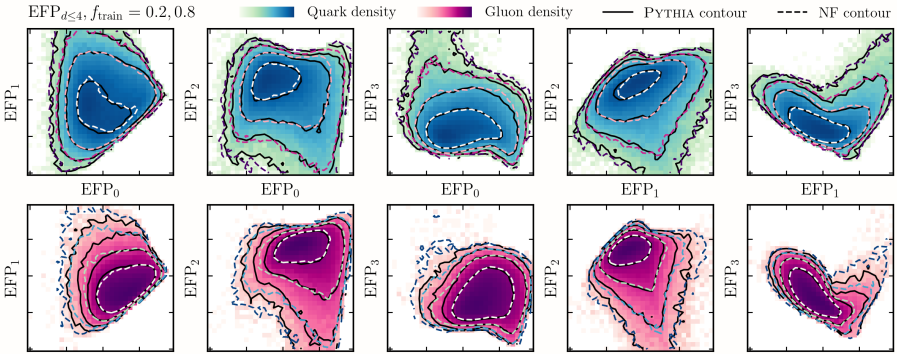
- Jets stored as prime Energy Flow Polynomials (EFPs) with
  - $1 \leq$ degree ≤ 4 (8 dimensional)
  - $1 \leq$ degree ≤ 6 (53 dimensional)
- Preprocess with PCA

Component distributions

- Good agreement with PYTHIA labels and topics.
- The first principal component is most challenging.
- Extrapolations are usually conservative.



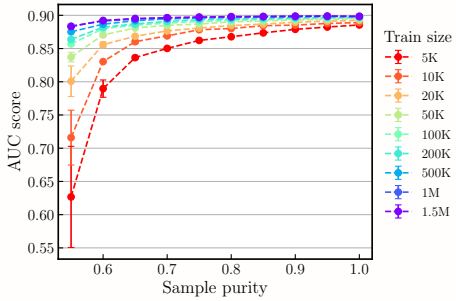
Component correlations

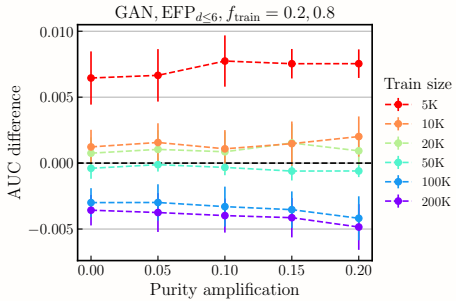


- Samples exhibit the correct correlations.
- Opens the possibility of combined use with other ML methods.

Dataset amplification

- Optimal weakly-supervised classifier requires maximizing size *and* purity of mixtures.
- Usually only one can be maximized.
- Conditional-generative networks can do both.
- Purity amplification improves performance on small datasets if initial purity is high.





Summary and Outlook

Summary

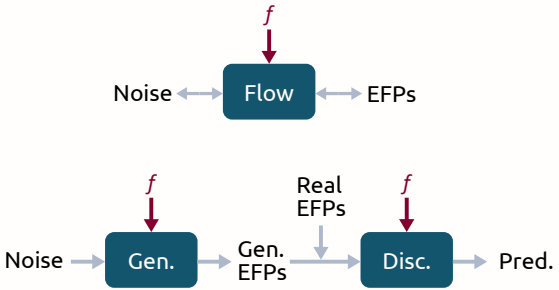
- Deep generative networks can complement histogram-based jet topics in high dimension.
- Simple purity-conditional approach gives good results.
- Purity amplification can benefit weakly-supervised classifier.

Outlook

- Reducibility-informed training (subtraction GAN, NF w. reverse-KL)
- CMS OpenData
- Jet images / particle clouds

Backup: Models and training

- **Normalising Flow**
  - Neural ODE Flow
  - 5×256 layers with skip connections
- **Generative Adversarial Network**
  - Wasserstein loss
  - Generator: 5×256 layers with skip connections
  - Critic: 5×256 layers
- Compatible with more than two mixtures.
- Simple conditioning (no prior enforcing compact purity).



Backup: Wasserstein distances

- Pure samples improve with training fraction.
- NFs scale with train size, while GANs are less sensitive.
- GANs perform similarly in high dimension.