Quark and Gluon Jet Discrimination

Matthew Schwartz
Harvard University

Based on work with J. Gallicchio arXiv:1211.7038, 1106.3076, 1104.1175
and P. Komiske and E. Metodiev arXiv:1612.01551

See also ATLAS arXiv:1405.6583, ATLAS-CONF-2016-034
Larkoski et al. arXiv:1405.6583
Why do we care?

1. BSM searches:
   New physics mostly **quark jets**
   Backgrounds mostly **gluon jets**

2. SM searches
   - Gluonic backgrounds to e.g. hadronic top decays

3. Improve Monte Carlos
   - Gluon jet modeling limits accuracy of current simulations

4. Test precision QCD

5. For the challenge: can we do it?
Quark/Glue basics

Probability of quark radiating:

\[ P(q \rightarrow qg) = \frac{\alpha_s}{2\pi} C_F (\cdots) \]

\[ C_F = \frac{4}{3} = 1.3 \]

Probability of quark radiating:

\[ P(g \rightarrow gg) = \frac{\alpha_s}{2\pi} C_A (\cdots) \]

\[ C_A = 3 \]

- Gluons around twice as likely to radiate than quarks
  - Gluon jets are fatter
  - Gluon jets are more massive
  - Gluon jets have more particles
  - ...

3
Example distributions

Jet mass

Charged particle count

Jet broadening

Jet angularity

$P_T$ fraction of 3$^{rd}$ hardest subjet

Moment of Hu
We looked at 10,000 variables

Integrated/differential “Jet Shape”

Jet angularities

Iteratively optimized radial profile

Properties of Covariance tensor

C = \sum_{i \in \text{jet}} \frac{p_T^i}{p_T^{\text{jet}}} \begin{pmatrix} \Delta \eta_i \Delta \eta_i & \Delta \eta_i \Delta \phi_i \\ \Delta \phi_i \Delta \eta_i & \Delta \phi_i \Delta \phi_i \end{pmatrix}

Combination of Eigenvalues

Eigenvalues: \( a > b \)

Quadratic Moment: \( g = \sqrt{a^2 + b^2} \)

Determinant: \( \text{det} = a \cdot b \)

Ratio: \( \rho = b/a \)

Eccentricity: \( \epsilon = \sqrt{a^2 - b^2} \)

Planar Flow: \( \text{pf} = \frac{4ab}{(a+b)^2} \)

Orientation: \( \theta \)

Subjet counts and properties

1st Subjet’s \( p_T \) Fraction

http://jets.physics.harvard.edu/qvg
We looked at 10,000 variables

The best two variables in Pythia are:

1. Charged particle count
   - Better spatial and energy resolution works better
     - e.g. particles > calorimeter clusters > subjets

2. Linear radial moment (girth)
   - Similar to jet broadening

\[ g = \frac{1}{p_T^{\text{jet}}} \sum_{i \in \text{jet}} p_T^i |r_i| \]

- Many variables have similar performance
Quark and gluon jet substructure

Significance Improvement

\[
\sigma = \frac{S}{\sqrt{B}} \quad \text{Cut} \quad \frac{S\epsilon_s}{\sqrt{B\epsilon_b}} = \sigma \frac{\epsilon_s}{\sqrt{\epsilon_b}}
\]

Top 5 combined with BDT

Girth

Particle count

Best group of 5
charged mult & girth
charged mult * girth
charged mult R=0.5
subjet mult R_{sub}=0.1
girth R=0.5
optimal kernel at 80%
1st subjet R=0.5
avg kT of R_{sub}=0.1
mass/Pt R=0.3
decluster kT R_{sub}=0.1
jet shape \Psi(0.1)
|pull| R=0.3
planar flow R=0.3

100 GeV jets, Pythia
ATLAS 7 TeV

- ATLAS developed procedure to disentangle quark and gluon jets
- Used relatively pure samples (dijets for gluon, $\gamma +$ jet for quark)

Extracted MC/MC

\[
\langle n_{trk}^j \rangle_{\text{MC/MC}} = 0.8, 1.0, 1.2
\]

Extracted Data/MC

\[
\langle n_{trk}^j \rangle_{\text{Data/MC}} = 0.8, 1.0, 1.2
\]

\[
\int \frac{d^2 \phi}{s} = 4.7 \text{ fb}^{-1}, \; \sqrt{s} = 7 \text{ TeV}
\]

\[
\frac{d^2 \phi}{s} = 0.4, \; |\eta| < 0.8
\]

Extracted gluon jet properties closer to quark than in pythia
Discrimination Performance

In order to determine which variables are most powerful for quark-gluon discrimination, a likelihood is created to rank the variables based on the fraction of gluons they reject (gluon rejection) for fixed quark sources, such as sample dependence, as discussed in Section 4.4 Validation.

The power of a variety of discriminants derived from data is studied and compared with predictions from Monte Carlo generators and high purity quark and gluon detector at the LHC. Templates of light-quark and gluon jet properties are derived and used to extract templates. The last bin in all plots includes overflow events.

ATLAS Preliminary

$\sqrt{s} = 8$ TeV 20.3 fb$^{-1}$

$1.2 < |\eta| < 2.1$

The ATLAS Collaboration

Data appears to be between Pythia and Herwig
Monte Carlo’s all seem to agree on quarks

- Promising sign for future progress

**Les Houches study (2016)**

**Quark, hadron-level**

- Pythia 8.205
- Herwig 2.7.1
- Sherpa 2.1.1
- Vincia 1.201
- Deductor 1.0.2
- Ariadne 5.0.β

**Gluon, hadron-level**

- Pythia 8.205
- Herwig 2.7.1
- Sherpa 2.1.1
- Vincia 1.201
- Deductor 1.0.2
- Ariadne 5.0.β

Monte Carlos can be improved...

Improved shower MCs in between Herwig and Pythia
Deep learning

Traditional approach

Think about physics

Guess some motivated observables

Run simulations

Combine best variables
  • Boosted Decision Trees
  • Neural Networks

Final multivariate discriminant

Test on data

Machine learning approach

Think about observations

Computer learns discrimination
Neural networks

Traditional (shallow) neural networks
Useful for multivariate analysis

Deep networks

- Many inputs
- Many hidden layers

**Activation function**

inspired by biology

\[
f(y) = \frac{1}{1 + \exp(-y)} \quad \text{(sigmoid)}
\]
Recent advances allowing Deep Learning

• In algorithms:
  – New activation functions to avoid issues such as saturation
  – New model regularizations
    • Dropout: Randomly selected fraction $p$ of units are ignored during each weight-update.
  – Network architecture adapted to application
    • Drastically fewer elements to optimize

• In computing:
  – Faster computing capabilities
    • Graphics Processing Units (GPUs)
  – Easier Usability
    • Keras Deep Learning Python Library
Activation Functions

- **Sigmoid**
  \[ f(y) = \frac{1}{1+\exp(-y)} \]

- **Tanh**
  \[ f(y) = \tanh(y) \]

- **Rectified Linear Unit (ReLU)**
  \[ f(y) = \max(0,y) \]

- Sigmoid and Tanh can **saturate**, whereby the gradient becomes vanishingly small for inputs far from zero, making training difficult.

- ReLUs avoid this saturation problem and have an easily computable gradient.
How to link up units?

- **Dense (Fully Connected) Layer**
  - Each unit is linked to every unit in the previous layer.

- **Convolutional Layer**
  - Each unit is linked to an $n \times n$ patch of the previous layer.
  - Units are downsampled to $n/2 \times n/2$ patches with a **max-pooling** layer.
  - Can handle multiple **channels**.
    E.g. RGB images.
Applications to Image Recognition
Jet Images  

Cogan et al. arXiv:1407.5675

- Treat energy deposits as image

**Application to boosted W tagging**

---

**Fisher's linear discriminant**

- Treat energy deposits as image

**preprocess**

**Fisher discriminant**

---

**Cell Coefficient**

---

**Figure 2:** A Fisher's linear discriminant presented as an image (left) and the distributions of the discriminant output when applied to W-jets and Light-jets (right), when the FLD is trained on jets with $p_T^2[250, 300]$ GeV, mass $M^2[65, 95]$ GeV, and separation between subjets of $R^2[0.6, 0.8]$.

The background rejection vs. signal efficiency curves for the FLD, computed using the 1-D likelihood ratios of the output distribution of the FLD for W-jets and QCD jets, can be seen in Figure 3a, along with the rejection vs. efficiency curves observed when using N-subjettiness ($\Delta^2 / \Delta^1$) [7, 8] computed analogously with the 1-D likelihood ratios. For the rejection vs. efficiency curve in Figure 3a Fisher-jets are trained on jets satisfying $p_T^2[250, 300]$ in 6 bins of $R_{jj}$, and a combined 1D likelihood ratio distribution is computed by taking the likelihood ratio for each jet computed with respect to appropriate $R_{jj}$ bin and merging these likelihood ratio values into a single distribution. The N-subjettiness distributions are not binned in $R_{jj}$ as this did not show any improvements in performance. Figure 3b shows the efficiency of W jets at a fixed QCD jet rejection of 10 as a function of jet $p_T$ for the FLD (combining the 6 bins of $R_{jj}$ for each jet $p_T$ bin) and for N-subjettiness. It can be seen that FLD outperforms N-subjettiness for the full range of jet $p_T$ examined.

It should be noted that the output of FLD and N-subjettiness are correlated, as shown in Figures 4a and 4b for W and QCD jets respectively, with a correlation coefficient of approximately 0.7 for both W and QCD jets. Thus, the Fisher-jet approach is able to combine in a linear way the information comprising the jet effectively, and capture much of the information of N-subjettiness and more. On the other hand, mass, which relies on quadratic relationships between the inputs, is a simple quantity which FLD does not reproduce, as shown in Figures 4c and 4d for W and QCD jets respectively. Since the Fisher-jet output is only slightly correlated with mass, with a correlation coefficient of approximately -0.25 for both W and QCD jets indicating a small degree of anti-correlation, the performance of the classifier does not change dramatically whether it is applied to a small window around the W mass, or to a sample without jet mass cuts.
Boosted W’s and jet images

de Olivera et al. arXiv:1511.05190

Baldi et al. arXiv:1603.09349

• Deep NN does better than single variables
• Deep NN does about as well as BDT
  of 6 good discriminants
Deep learning for Q vs G

- Find anti-$k_T$ $R=0.4$ jets
- Extract square grid around jet center
- Pixelate into $\Delta \eta \times \Delta \phi = 0.024 \times 0.024$ cells
  - Produces 33x33 image

**We use 3 “colors”**

- Red = $p_T$ of charged particles
- Green = $p_T$ of neutral particles
- Blue = charged particle multiplicity

**preprocess**
- Center
- Crop
- Normalize
- Zero
- Standardize

**NN inputs**
Deep NN architecture

- Convolution layers apply 8x8 pixel filters to images
  - 4x4 filters used for for 2\textsuperscript{nd} and 3\textsuperscript{rd} conv. layers
  - We use 64 independent filters in each layer
- Max-pooling reduces layer size by 4
- Final layer is densely connected to all final filters

Figure 2: An illustration of the deep convolutional neural network architecture. The first layer is the input jet image, followed by three convolutional layers, a dense layer and an output layer.
Results

The ROC and SIC curves of the jet variables and the deep convolutional network on 200 GeV and 1000 GeV Pythia jets are shown in Figure 5. The quark jet classification efficiency at 50% quark jet classification efficiency for each of the jet variables and the CNN are listed in Table 1. To combine the jet variables into more sophisticated discriminants, a boosted decision tree (BDT) is implemented with scikit-learn. The convolutional network outperforms the traditional variables and matches or exceeds the performance of the BDT of all of the jet variables. The performance of the networks trained on images with and without color is shown in Figure 6.

Works really well – especially considering we don’t put in any physics!
Results

The ROC and SIC curves of the jet variables and the deep convolutional network on 200 GeV and 1000 GeV Pythia jets are shown in Figure 5. The quark jet classification efficiency at 50% quark jet classification efficiency for each of the jet variables and the CNN are listed in Table 1. To combine the jet variables into more sophisticated discriminants, a boosted decision tree (BDT) is implemented with scikit-learn. The convolutional network outperforms the traditional variables and matches or exceeds the performance of the BDT of all of the jet variables. The performance of the networks trained on images with and without color is shown in Figure 6.

Works really well – especially considering we don’t put in any physics!
Comparing Pythia and Herwig

- Discrimination worse in Herwig
  - Gluon and quark jets are more similar
  - Consistent with previous studies

Network performance independent of MC used to train

- Indicates robustness
- May work on data
Is it learning physics?

Add in observables
- CPM = charged particle multiplicity
- $N_{95}$ = a useful discriminant (minimum number of pixels with 95% of jet $p_T$) [Pumplin 1991]

Except at very high $p_T$, no benefit from adding observables
May indicate that NN has “learned” physics
Conclusions

• Quark and gluon jets can be distinguished by radiation patterns
  • Pythia and Herwig have significant differences, particularly for gluons
  • Improved parton showers (e.g. vincia) look promising

• Traditional variables
  • Two types: shape (mass, girth, n-jettiness) and count (# particles, # subjets)
  • Marginal gains from exploiting correlations of >2 variables using BDTs

• Deep learning approach
  • Use image-recognition technology to avoid thinking
  • Does better than traditional approach!
  • Relies heavily on simulations, but
    • Performance independent of Pythia or Herwig training

Domo Arigato!
BACKUP
Analytic approach to correlations

Pythia 8
$I(T;\lambda^k_\beta)$
quark vs. gluon
$p_T > 400$ GeV
$R_0 = 0.6$

Herwig++
$I(T;\lambda^k_\beta)$
quark vs. gluon
$p_T > 400$ GeV
$R_0 = 0.6$

Next-to-LL
$I(T;\lambda^k_\beta)$
quark vs. gluon
$p_T > 400$ GeV
$R_0 = 0.6$

Analytic approach to generalized angularities

\[ \lambda^k_\beta = \sum_{i\in\text{jet}} z_i^k \theta_i^\beta . \]

- Challenging
- Not impossible
- Complementary to MCs
Data: where are the quark and gluon jets?

200 GeV Quark Purity (zoom)

200 GeV Gluon Purity

Photon + jet samples
• Jet closer to photon likely quark

• b + 2 jet: high purity low s
  • One jet is b other is gluon
• Dijet: high cross section, low purity
Multivariate approach

- We can think about and visualize **single variables**

- Two variables are harder

- Multidimensional distributions are not well-suited for visualization.

- There are things that **computers are just better** at.

- Multivariate approaches let you figure out how well you could **possibly do**

**FRAMING**
See if simple variables can do as well (establishes the goal)

**EFFICIENCY**
Save you the trouble or looking for good variables (project killer)

**POWER**
Sometimes they are really necessary (e.g. b tagging)
Multivariate methods

Lots of methods

• Boosted Decision Trees
• Artificial Neural Networks
• Fischer Discriminants
• Rectangular cut optimization
• Projective Likelihood Estimator
• H-matrix discriminant
• Predictive learning/Rule ensemble
• Support Vector Machines
• K-nearest neighbor
• ...

Useful in many areas of science

For particle physics, **Boosted Decision Trees** are best suited for combining variables

- Easy to understand
- Train fast
- Nearly optimal efficiencies
Are they correlated?

- Not completely
- Can get more discrimination from 2D cuts
• Pythia and Herwig qualitatively similar
• Discrimination power with Herwig ++ universally worse
# Quark and gluon tagging: results

## Gluon Efficiency % at 50% Quark Acceptance

<table>
<thead>
<tr>
<th>Gluon Efficiency % at 50% Quark Acceptance</th>
<th>50 GeV</th>
<th>200 GeV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Particles</td>
<td>Tracks</td>
</tr>
<tr>
<td></td>
<td>P8</td>
<td>H++</td>
</tr>
<tr>
<td>2-Point Moment $\beta=1/5$</td>
<td>8.7*</td>
<td>17.8*</td>
</tr>
<tr>
<td>1-Subjettiness $\beta=1/2$</td>
<td>9.3</td>
<td>18.5</td>
</tr>
<tr>
<td>2-Subjettiness $\beta=1/2$</td>
<td>9.2</td>
<td>18.6</td>
</tr>
<tr>
<td>3-Subjettiness $\beta=1$</td>
<td>9.1</td>
<td>19.3</td>
</tr>
<tr>
<td>Radial Moment $\beta=1$ (Girth)</td>
<td>10.3</td>
<td>20.5</td>
</tr>
<tr>
<td>Angularity $a=+1$</td>
<td>10.3</td>
<td>20.0</td>
</tr>
<tr>
<td>Det of Covariance Matrix</td>
<td>11.2</td>
<td>21.2</td>
</tr>
<tr>
<td>Track Spread: $\sqrt{\langle p_T^2 \rangle/p_T^{jet}}$</td>
<td>16.5</td>
<td>25.3</td>
</tr>
<tr>
<td>Track Count</td>
<td>17.7</td>
<td>26.4</td>
</tr>
<tr>
<td>Decluster with $k_T$, $\Delta R$</td>
<td>15.8</td>
<td>24.5</td>
</tr>
<tr>
<td>Jet $m/p_T$ for R=0.3 subjet</td>
<td>13.1</td>
<td>25.9</td>
</tr>
<tr>
<td>Planar Flow</td>
<td>28.7</td>
<td>34.4</td>
</tr>
<tr>
<td>Pull Magnitude</td>
<td>37.0</td>
<td>39.0</td>
</tr>
<tr>
<td>Track Count &amp; Girth</td>
<td>9.9</td>
<td>20.1</td>
</tr>
<tr>
<td>R=0.3 $m/p_T$ &amp; R=0.7 2-Point $\beta=1/5$</td>
<td>7.9*</td>
<td>17.7</td>
</tr>
<tr>
<td>1-Subj $\beta=1/2$ &amp; R=0.7 2-Point $\beta=1/5$</td>
<td>8.5</td>
<td>17.3*</td>
</tr>
<tr>
<td>Girth &amp; R=0.7 2-Point $\beta=1/10$</td>
<td>12.6</td>
<td>21.9</td>
</tr>
<tr>
<td>1-Subj $\beta=1/2$ &amp; 3-Subj $\beta=1$</td>
<td>8.9</td>
<td>18.0</td>
</tr>
<tr>
<td>Best Group of 3</td>
<td>7.5</td>
<td>17.0</td>
</tr>
<tr>
<td>Best Group of 4</td>
<td>7.1</td>
<td>16.7</td>
</tr>
<tr>
<td>Best Group of 5</td>
<td>6.9</td>
<td>16.4</td>
</tr>
</tbody>
</table>

**Table 1.** Comparison of gluon efficiency at the 50% quark acceptance working point. All of the single variables use R=0.5 jets, whereas combinations sometimes include R=0.7 jets. Gluon efficiencies, rather than gluon rejections (one minus efficiencies), are shown because a fractional improvement here is the same fractional improvement in $S/B$. Divided by two, it is also the fractional improvement in $S/\sqrt{B}$. These scores have ±0.5% statistical errors, but they are correlated — the differences between variables has smaller spread, as does the improvement when combining variables. Because of the large number of variables and parameters, and the larger number of possible combinations of these, there is definitely a look-elsewhere-type effect when choosing the top pair. Many pairs statistically tied for the top spot in each category, so five pairs were chosen as representative. Their scores are marked with asterisks, as are the best individual variables in each category. These group of 3, 4, and 5 starts to show diminishing returns.