Boosted W tagging with Lund jet planes

Jad M. Sardain on behalf of the ATLAS Collaboration

First Lund Jet Plane Institute, July 3-7 2023
- Heavy states with large momentum will produce multi-prong topology

- Techniques (ML or not) developed in the last 10 years to distinguish the various jet topologies

- Various approaches include:  
  - Combine various high-level variables
  - Use low-level constituents
  - Jet images
  - Jet clustering sequence (this talk)
Jet tagging using Lund plane and ML

- The Lund planes for the W, top and QCD jets already looks quite different
  - Convolutional NN used for identification
  - Results not much better than CNN directly on jet images

- However, the Lund plane has much more information coming from the sequence that produced it
  - Use GNN where each node has 3 vars:
    - $z$ momentum fraction of the branching.
    - $k_t$ transverse momentum,
    - $\Delta$ emission angle
  - Number of tracks per jet as a global feature to help classification

**Signal**

**Background**

W-jet

QCD
LundNet

- Graph neural network by Frédéric Dreyer used to tag Lund planes when represented as graphs. It is currently the state of the art of Lund tagging and inspired by ParticleNet \[\text{arXiv:2012.08526v2 [hep-ph]}\]

- It uses the EdgeConv layer. In summary, for a node \(x_i\), we construct a small fully connected neural network. The input is \(x_j - x_i = [k_{t,j} - k_{t,i}, \Delta_j - \Delta_i, z_j - z_i]^T\) concatenated with \(x_i\), where \(x_j\) is just a node connected to \(x_i\). The output is the edge features, \(e\).

- The EdgeConv block repeats this operation for every node connected to \(x_i\). Then the edge features are aggregated (based on taking the mean) to produce the new node features for \(x_i\).

- LundNet-3 and LundNet-5 are virtually the same model, their difference is the number of Lund variables each node has at the beginning. In our analysis, we only consider LundNet-3.
Signal: $W' \rightarrow WZ$
- Truth matched to $W$
- $m(\text{truth jet}) > 50$ GeV
- Number of $b$-hadrons $= 0$

Background: QCD
- Jet is truth-matched
- $p_T(\text{truth jet}) > 200$ GeV
- $200 < p_T(\text{jet}) < 3000$ GeV, $|\eta(\text{jet})| < 2.0$
- $40 < m(\text{jet}) < 300$ GeV

- Outline of the analysis: 1) Classifier 2) pre-train adversarial 3) combined train for mass decorrelation

$$\mathcal{L} = w_{clf} \cdot \sum_{i \in (s+b)} L_{\text{classifier}} + w_{adv} \cdot \lambda \cdot \sum_{i \in b} L_{\text{decor}}$$

- The adversarial network is a gaussian mixture model that use 20 gaussians to infer the correlation between the output score of the classifier and the mass

For each Gaussian of 20:
- $\mu$ : mean
- $\sigma$ : std.
- $\pi$ : norm
Baseline taggers

- Two taggers are considered as baseline taggers and are used for comparisons:
  - The “so-called” 3-var tagger, based on number of tracks, mass and D2
  - The DNN/ANN tagger, based on high level observables
Baseline taggers

- Two taggers are considered as baseline taggers and are used for comparisons:
  - The “so-called” 3-var tagger, based on number of tracks, mass and D2
  - The DNN/ANN tagger, based on high level observables

- Cuts tuned in order to achieve a certain signal efficiency \( \varepsilon^\text{sig} = \frac{N^{\text{tagged}}}{N^{\text{total}}_\text{sig}} \)
  (Example show here for a WP@50%)

- An jet is tagged if:
  \[
  m_{\text{low}}^\text{cut} < m < m_{\text{high}}^\text{cut} \\
  D_2 < D_2^\text{cut} \\
  N_{\text{trk}} < N_{\text{trk}}^\text{cut}
  \]
Baseline taggers

- Two taggers are considered as baseline taggers and are used for comparisons:
  - The “so-called” 3-var tagger, based on number of tracks, mass and D2
  - The DNN/ANN tagger, based on high level observables

- High level observables are used as input features for the classifier

- Same strategy as LundNet, classifier then adversarial

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_2$, $C_2$</td>
<td>Energy correlation ratios</td>
</tr>
<tr>
<td>$\tau_{21}$</td>
<td>N-subjettiness</td>
</tr>
<tr>
<td>$R^F_2$</td>
<td>Fox-Wolfram moment</td>
</tr>
<tr>
<td>$\mathcal{P}$</td>
<td>Planar flow</td>
</tr>
<tr>
<td>$a_3$</td>
<td>Angularity</td>
</tr>
<tr>
<td>$A$</td>
<td>Aplanarity</td>
</tr>
<tr>
<td>$Z_{cut}$, $\sqrt{d_{12}}$</td>
<td>Splitting scales</td>
</tr>
<tr>
<td>$K\tau\Delta R$</td>
<td>$k_\tau$-subjet $\Delta R$</td>
</tr>
</tbody>
</table>
Baseline taggers

- Two taggers are considered as baseline taggers and are used for comparisons:
  - The “so-called” 3-var tagger, based on number of tracks, mass and D2
  - The DNN/ANN tagger, based on high level observables

To decorrelate jet mass and DNN score:
- Apply an additional adversarial neural network (ANN) to the DNN tagger

- ANN trained to infer the jet mass from the DNN score by minimizing $L_{ANN}$

- Loss function of the combined training $L_{total} = L_{DNN} - \lambda L_{ANN}$, with $\lambda$ being chosen with a compromise between the background rejection and the mass decorrelation
Baseline taggers

- Two taggers are considered as baseline taggers and are used for comparisons:
  - The “so-called” 3-var tagger, based on number of tracks, mass and D2
  - The DNN/ANN tagger, based on high level observables

Background rejection rate comparison of W taggers

- DNN tagger (violet solid) shows the best performance
- Decrease in performance after ANN is expected

---

**Figure 1:**

*ATLAS* Simulation Preliminary

\( \sqrt{s} = 13 \) TeV, W jet tagging

- anti-\( k_t \), R=1.0 UFO Soft-Drop CS+SK jets
- \( p_T \in [500, 1000] \) GeV
- Cut on \( m_J \) from 3-var tagger

Background rejection rate 1/\( \epsilon_{\text{sig}} \)

<table>
<thead>
<tr>
<th>Analytical</th>
<th>MVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D_2 )</td>
<td>( Z_{\text{NN}} )</td>
</tr>
<tr>
<td>( D_2^{\text{NN}} )</td>
<td>( Z_{\text{ANN}}^{(x=10)} )</td>
</tr>
<tr>
<td>3-var</td>
<td></td>
</tr>
</tbody>
</table>

Background rejection rate 1/\( \epsilon_{\text{sig}} \)

<table>
<thead>
<tr>
<th>Analytical</th>
<th>MVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>( D_2^{\text{NN}}/D_2 )</td>
<td>( Z_{\text{ANN}}^{(x=10)}/Z_{\text{NN}} )</td>
</tr>
<tr>
<td>( D_2^{\text{NN}}/D_2 )</td>
<td></td>
</tr>
<tr>
<td>( Z_{\text{ANN}}^{(x=10)}/Z_{\text{NN}} )</td>
<td></td>
</tr>
</tbody>
</table>

Signal efficiency \( \epsilon_{\text{rel}} \)

<table>
<thead>
<tr>
<th>MVA</th>
<th>Analytical</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Z_{\text{ANN}}^{(x=10)}/D_2^{\text{NN}} )</td>
<td></td>
</tr>
<tr>
<td>( Z_{\text{ANN}}^{(x=10)}/D_2^{\text{NN}} )</td>
<td></td>
</tr>
<tr>
<td>( Z_{\text{NN}}/D_2 )</td>
<td></td>
</tr>
</tbody>
</table>

Signal efficiency \( \epsilon_{\text{rel}} \)
Baseline taggers

- Two taggers are considered as baseline taggers and are used for comparisons:
  - The “so-called” 3-var tagger, based on number of tracks, mass and D2
  - The DNN/ANN tagger, based on high level observables

- Adding the information on the number of tracks helped in increasing the background rejection

- Previously, the 3-variable tagger showed better performance than DNN

- Now, DNN performing better than the 3-variable cut based tagger, and the ANN is comparable with the 3-variable tagger performance

- Reason: Most other feature exploit the 2 prong behavior of the W/Z decay, whereas the number of tracks is a good quark/gluon discriminator
Let's go back to LundNet

- The LundNet tagger without mass decorrelation achieves the best performance

- The adversarial network significantly deteriorates performance (for both LundNet and DNN)

- At 50% signal efficiency and with the $p_T$-dependent 3-var tagger mass cut, the background rejection, after mass decorrelation, is better by a factor of 2.5(3) with respect to the 3-var tagger (baseline ANN tagger)
- The LundNet tagger without mass decorrelation achieves the best performance

- The adversarial network significantly deteriorates performance (for both LundNet and DNN)

- At 50% signal efficiency and with the $p_T$-dependent 3-var tagger mass cut, the background rejection, after mass decorrelation, is better by a factor of 2.5(3) with respect to the 3-var tagger (baseline ANN tagger)
- Across all $p_T$ ranges:
  - LundNet$^{NN}$ is able to retrieve a peak around the W mass
  - LundNet$^{ANN}$ is able to retrieve the shape of the QCD background

- To quantify the agreement, the KL divergence was calculated:
  - Got values < 1% for both comparison:
    LundNet$^{NN}$ with signal
    LundNet$^{ANN}$ with QCD

Backup: results for WP@80%
- The LundNet tagger shows a decrease in background rejection of 20%-40% for 50% working point.

- Higher contribution in the region factorizing the hard collinear emission for Sherpa with string model than Sherpa using the cluster model.

- Herwig with angle ordered parton shower has a higher contribution from soft collinear emission than Herwig with dipole parton shower.
- Jets are not just an image, they are a process that can be measured by deconstructing the jet clustering algorithm

- This is the ideal field of applications of a GNN

- Results are better than other methods, but mass sculpting shows up in background peaking at $m(W)$

- Use of adversarial network solves the issue but reduces performance

- Good mass decorrelation and background rejection in all $p_T$ intervals

- Mass correlated tagger tests using other MC generators result in good background rejection
Backup
Results for WP@80%

**ATLAS Simulation Preliminary**

\( \sqrt{s} = 13 \text{ TeV}, W \) tagging

anti-\( k_t \), \( R=1.0 \) UFO Soft-Drop CS+SK jets

\( \varepsilon_{\text{sig}} = 80\% \)

Cut on \( m_\gamma \) from 3-var tagger

---

**ATLAS Simulation Preliminary**

\( \sqrt{s} = 13 \text{ TeV}, W \) tagging

anti-\( k_t \), \( R=1.0 \) UFO Soft-Drop CS+SK jets

\( \varepsilon_{\text{sig}} = 80\% \; p_T \in [200, 500] \text{ GeV} \)

---

**ATLAS Simulation Preliminary**

\( \sqrt{s} = 13 \text{ TeV}, W \) tagging

anti-\( k_t \), \( R=1.0 \) UFO Soft-Drop CS+SK jets

\( \varepsilon_{\text{sig}} = 80\% \; p_T \in [500, 1000] \text{ GeV} \)

---

**ATLAS Simulation Preliminary**

\( \sqrt{s} = 13 \text{ TeV}, W \) tagging

anti-\( k_t \), \( R=1.0 \) UFO Soft-Drop CS+SK jets

\( \varepsilon_{\text{sig}} = 80\% \; p_T \in [1000, 3000] \text{ GeV} \)
ATLAS Simulation Preliminary

\( \sqrt{s} = 13 \text{ TeV, } W \text{ tagging} \)

anti-\(k_t\), \(R=1.0\) UFO Soft-Drop CS+SK jets

\( \epsilon^{\text{rel}}_{\text{sig}} = 80\% \)

LundNet

\( \text{Background rejection } 1/\epsilon^{\text{rel}}_{\text{bkg}} \)

\( \text{Alternative / Pythia} \)

\( \text{Large-}\(R\) jet \(p_T\) [GeV]}

\( \text{Pythia} \)

\( \text{Sherpa Lund} \)

\( \text{Sherpa Cluster} \)

\( \text{Herwig Angular} \)

\( \text{Herwig Dipole} \)