Deep Learning Jet Substructure from Two Particle Correlation

Yang-Ting Chien

YITP-CFNS Fellow, Stony Brook University

In collaboration with Kai-Feng Chen (National Taiwan University)
arXiv:1911.02020, and work in progress

Stony Brook University

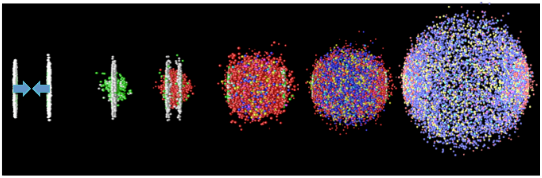
ML4Jets
NYU, January 15, 2020

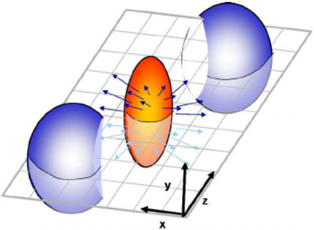
Center for Frontiers in Nuclear Science

Outline

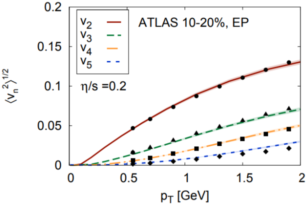
- Two-particle correlation as jet representation
  - fundamental information unit of particle relations
- Correlate with physics analysis
  - telescoping deconstruction: an expansion of subjet observables
  - soft-drop and collinear-drop
- Conclusion and outlook

Heavy ion collisions and quark gluon plasma





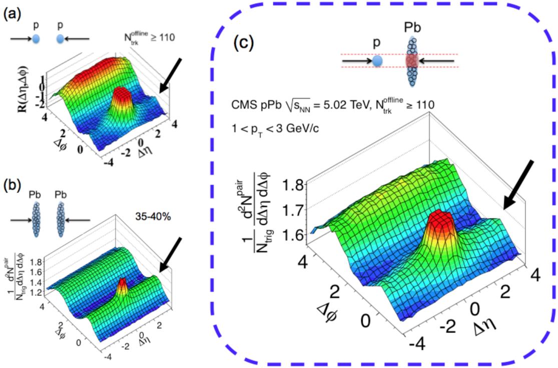
- A hot and dense medium is created
  - The medium quickly thermalizes and evolves into $\mathcal{O}(10^5)$ soft hadrons
  - Soft particle distributions described well with
    - Geometric and fluctuating initial stages
    - Hydrodynamics and small values of η/s
  - QGP: a droplet of perfect liquid?
- Sometimes energetic jets are also produced within the medium simultaneously



$$\frac{dN}{d\phi} = \sum_n v_n \cos n(\phi - \phi_n)\ ,\ \phi : \text{azimuth}$$

Long range correlation $\Delta\phi \approx 0$, $\Delta\eta$ large

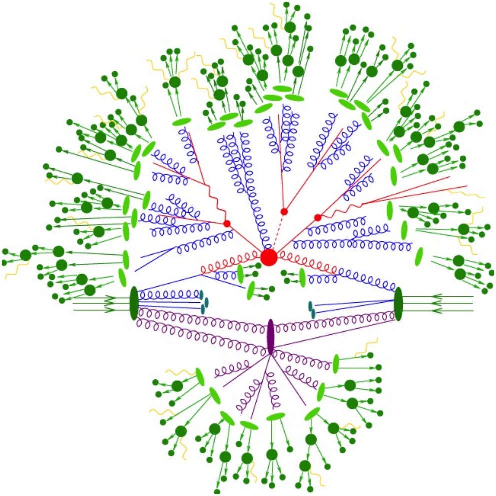
CMS, Phys. Lett. B 718 (2013) 795, JHEP 09 (2010) 091, JHEP07(2011)076



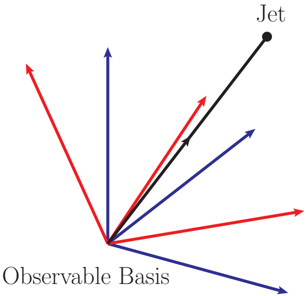
- A signature of QGP seen in two particle correlation in pp, pA and AA collisions
- The smallest droplet of liquid? What do "standard" pp simulations say about this?

Challenge and opportunity in nuclear and particle physics simulations

- pp event simulation paradigm
  - parton shower
  - underlying events
  - hadronization
- Burning issues
  - quark-gluon plasma signature in pp, pA and AA collisions
  - hydrodynamics and collectivity
  - understanding initial state dependence is essential
- Concrete strategy to study any stage of collider event $\equiv$ jet substructure
- Can machine learning help?



Jet representations



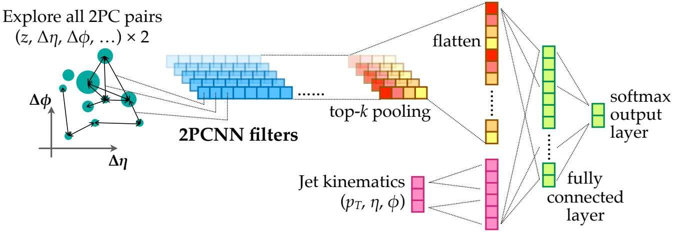
- Different multivariate techniques/machine learning architectures suit different jet representations, and vice versa
  - list of physics-motivated observables (conventional)
  - unbiased, raw input (particle momenta, PID, image, tree, graph, point cloud, ...)
  - complete basis and expansion (Nsubjettiness, EFP/EFN, telescoping deconstruction, ...)
- The rise of machine learning gives powerful tools for extracting physics features

- Use two-particle correlation (2PC): pairs of particle i and j as input jet representation
  - $C_2^N \propto N^2 \gg N$, a redundancy of jet information
  - Help efficiently build up jet features which can be probed with concrete observabes
- Illustrate using supervised learning in a variety of classification tasks

Tasks, samples, and inputs

- We explore a few tasks which exploit qualitatively different features
  - two-prong tagging: W versus light quark
  - two-prong tagging + vertex: Higgs$\to b\bar{b}$ versus light quark
  - three-prong tagging: top versus light quark
  - W^+ versus W^-: electric charge (inspired by David's work)
  - quark versus gluon: color and flavor
- Samples are generated from MC simulations using MadGraph and Pythia 8 and reconstructed as anti-kT $R = 0.8$ ($R = 0.4$ for quark/gluon discrimination) jets
  - $Z' \to W^+W^-, ZH, t\bar{t}, q\bar{q}$, same hard kinematics
  - $m_{Z'} = 2$ TeV
  - QCD for quark and gluon jets
- Truth particle information is passed through a Delphes simulation into track, Ecal and Hcal information
- 2PC Inputs: $z = p_T^i/p_T(\text{jet})$, $\Delta\eta = \eta^i - \eta(\text{jet})$, $\Delta\phi = \phi^i - \phi(\text{jet})$ + rotation (preprocessing)
  - The basic input layer consists of energy flow information
  - An extra layer consists of track information (charge and 2PC vertex)

Two-particle correlation neural network (2PCNN) using Keras + TensorFlow



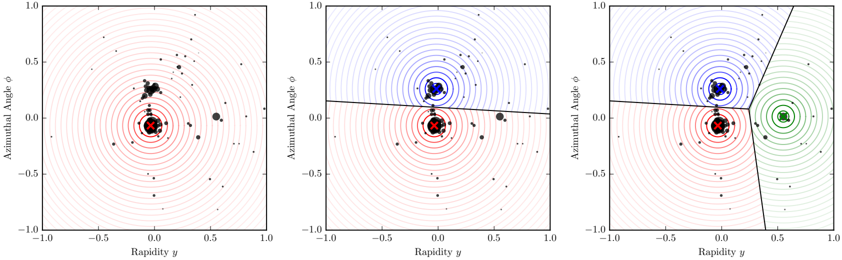
- Use a collection of filters (64, 32 for the track layer) with shared weight to process 2PCs
  - Each filter is a fully connected dense network which gives outputs to all the 2PCs
  - Only top-k (e.g. k=4) ranked 2PCs are kept as inputs for the subsequent decision-making, fully connected network
  - Analogy: ants (filters) going out to find food (2PC features)
- Baseline jet kinematic information is included with a dense network
- Outputs of 2PCNN layer and dense network are followed by a fully connected layer (128 nodes, ReLU) and two output nodes (softmax)
- We use cross-entropy loss function and Adam optimizer
- Details in example code and test sample available at https://github.com/kfjack/2PCNN

Some words on the comparison with other methods

- Particle Cloud with ParticleNet (1902.08570)
  - similarity: treating particle inputs as sets and using correlations
  - difference: 2PCNN does not use convolution while ParticleNet uses edge convolution
- Energy Flow Network (1810.05165) and spectral analysis (Sung Hak's talk)
  - similarity: building upon particle correlation
  - difference: 2PCNN stays at the level of 2PCs while EFN/SA treat observables
- Convolutional neural network
  - similarity: using filters
  - difference: at the input level 2PCNN filters are global while CNN filters are local
- In order to benchmark the 2PCNN performance, we compare with telescoping deconstruction

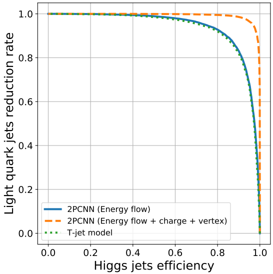
Telescoping Deconstruction: a complete subjet expansion

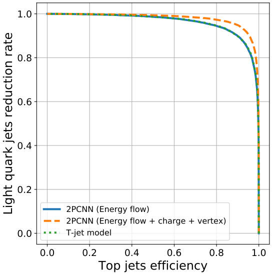
- A fast converging, fixed-order N subjet expansion with subjet kinematics information
  - identify dominant energy flow directions using N soft recoil-free axes
  - reconstruct subjets around the axes with multiple subjet radii R
  - TD variables respects the IR structure of QCD when organizing information
- Closely related to perturbative expansion and parton shower picture
- Truncate at $N = 3$ with four radius values. Totally 60 input variables to the previous, same dense network (128 nodes, ReLU). Fast and powerful.



Chien, Elayavalli, 1803.03589

ROC curves for Higgs and top tagging





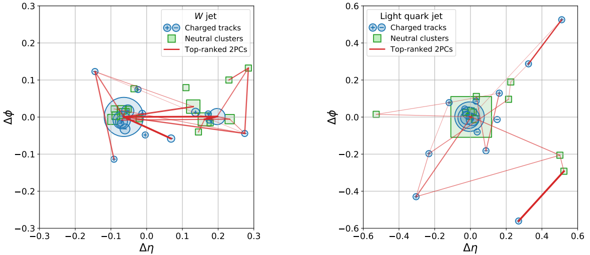
- Performance based on energy flow information is comparable to or higher than TD
  - A consistency check and a benchmark of 2PCNN performance
- Vertex information is useful because of the secondary b vertex in Higgs$\rightarrow b\bar{b}$ and $t \rightarrow W + b$

Performance overview

	2PCNN(E-flow)		2PCNN(full)		T-jet model	
Task	ACC	AUC	ACC	AUC	ACC	AUC
W vs quark	0.881	0.945	0.881	0.946	0.880	0.945
Higgs vs quark	0.873	0.939	0.959	0.993	0.866	0.934
top vs quark	0.900	0.962	0.929	0.978	0.900	0.963
W^+ vs W^-	0.505	0.502	0.757	0.839	0.502	0.502
quark vs gluon	0.738	0.810	0.748	0.823	0.732	0.802

- The practical: excellent classification performance and feature extraction quantified by AUC (area under ROC curves) and ACC (accuracy)

Illuminate trained models with filter outputs

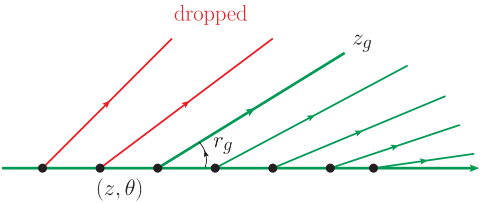


- The importance of the top-k ranked 2PC pairs within a filter can potentially be quantified by their filter output values 2PCNN has learned
  - Top-one ranked 2PC pair of each active filter is indicated by a solid line, with the thickness representing the strength of the filter output
- Jet constituents: scattered circles and squares, sizes $\propto$ particle transverse momenta
- Two distinct features
  - correlations within and between the prongs
  - correlations between high pT constituents within the prongs and low pT constituents scattered at wide angle

Neural network correlated with physical analysis

Dasgupta, Fregoso, Marzani, Salam, JHEP09(2013)029

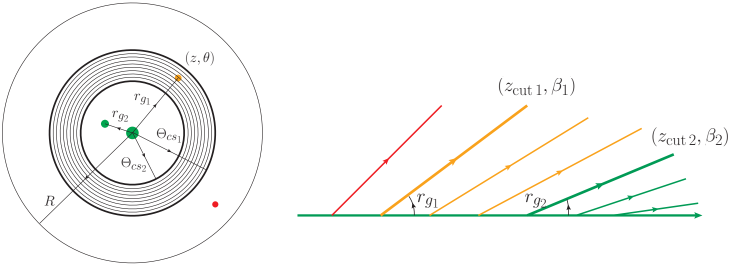
Larkoski, Marzani, Soyez, Thaler, JHEP05(2014)146



- Soft Drop: tree-based procedure to drop soft radiation
  - Recluster a jet using Cambridge-Aachen algorithm into an angular-ordered tree
  - For each branching, consider the p_T of each branch and the angle θ between branches
  - Soft drop condition: drop the soft branch if $z < z_{\text{cut}}\ (\theta/R)^{\beta}$, where z is the momentum fraction of the soft branch
  - We use $z_{\text{cut}} = 0.2$ and $\beta = 0$

Collinear Drop using soft drop + anti soft drop

Chien, Stewart, 1907.11107

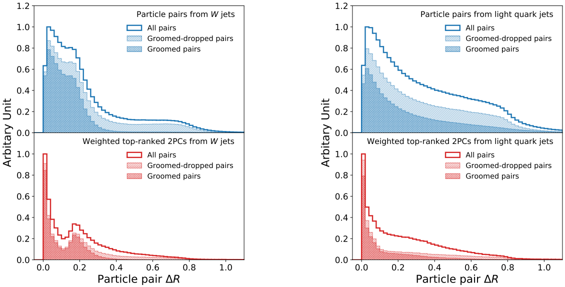


- Probe the soft radiation within the ring characterized by energies E_{cs_i} and angles Θ_{cs_i}
- Phase space constraints on soft emissions with $(\text{z}, \theta) = (\text{momentum fraction}, \text{angle})$,

$$z_{\text{cut}\,1}\left(\frac{\theta}{R}\right)^{\beta_1} \lesssim z \lesssim z_{\text{cut}\,2}\left(\frac{\theta}{R}\right)^{\beta_2}$$

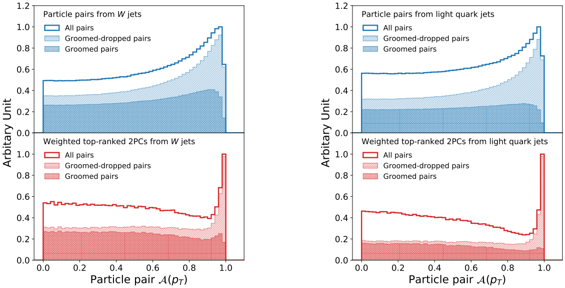
- Classify jet constituents into groomed and dropped categories
  - 2PCs form distinct sets: groomed-groomed, groomed-dropped and dropped-dropped

2PC angular correlation $\Delta R = \sqrt{(\eta^i - \eta^j)^2 + (\phi^i - \phi^j)^2}$ distribution



- To maximize the sensitivity to extracted features, lower panels show the top-ranked 2PC distributions weighed by the output values of 2PCNN filters
  - For W jets, strong features are identified at $\Delta R \approx 0$ and $\Delta R \approx 0.2 \sim 2m_W/p_T(\text{jet})$
  - For light quark jets the $\Delta R \approx 0$ feature is strong and the $\Delta R \approx 0.2$ feature is absent
- One and two-prong structures are dominantly determined by the groomed-groomed 2PC pairs
- Upper panels show corresponding distributions with equal weight

2PC p_T asymmetry $\mathcal{A} = |p_T^i - p_T^j|/(p_T^i + p_T^j)$ distribution



- lower panels show the top-ranked 2PC distributions weighed by the output values of 2PCNN filters
  - a clear feature at $\mathcal{A} \approx 1$ in distributions for both samples
- The feature at $\mathcal{A} \approx 1$ dominantly comes from the groomed-dropped 2PC pairs which correlate hard, collinear particles to soft, wide-angle particles: color-singlet isolation

Conclusion and outlook

- We construct a new two-particle correlation neural network
- 2PCNN with energy flow information perform comparably as telescoping deconstruction
- 2PCNN can easily include charge and vertex information with significant improvement
- Filter outputs can be directly extracted and used to illuminate trained network
- Extensions to new tasks and event level studies are straighforward
- Check out https://github.com/kfjack/2PCNN!

Public repository for 2PCNN

7 commits | 1 branch | 0 packages | 0 releases | 1 contributor

Branch: master | New pull request | Find file | Clone or download

kfjack Update README.md — Latest commit 6f889e5 6 days ago

README.md	Update README.md	6 days ago
prototype_deploy.py	Add files via upload	7 days ago
prototype_train.py	Add files via upload	7 days ago
wgts_2pcnn_fatjet_t_vs_q.h5	Add files via upload	7 days ago
wgts_2pcnn_fatjet_w_vs_q.h5	Add files via upload	7 days ago