Metric Methods with Open Collider Data
Patrick T. Komiske, Radha Mastandrea, Eric M. Metodiev, Preksha Naik, Jesse Thaler
Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Introduction
Machine learning and particle physics are on a collision course, producing exciting new ideas. Exploring public collider data inspired new fundamental questions, with answers coming from an unlikely place: optimal transport.

Q: What's the "distance" between two collisions? A: The "work" to rearrange one into another!

The "work" to rearrange one event into another! Equipping collider data with a metric unlocks new unsupervised and visualization techniques. Remarkably, this connection sheds new light on fundamental questions, with answers coming from an unlikely place: optimal transport.

A Metric for Collider Data

When are two particle collisions similar? Or when are two jets similar?

Natural question with no satisfying answer in physics. Image-based pixel comparisons are unstable under tiny perturbations of the particles. Observable (i.e. feature) comparisons can have zero distance for very different events or jets.

The Earth (or Energy) Mover’s Distance (EMD) provides a natural answer. Solving for the EMD is an optimal transport problem. The "work" to rearrange one event into another!

Collider Physics and Optimal Transport

Six decades of collider techniques can naturally cast as geometry in the "space of events" with EMD.

Smooth functions of energy distribution are finite in QFT. Jets are N-particle event approximations. Jets are "more than fractal". Subtract pileup as a uniform distribution.

Exploring the Space of Jets

The "space of jets" can be visualized by embedding the jet dataset with t-SNE.

25-medoid jets shown, sized by importance. 25-medoid jets shown, sized by importance. A peak of one-prong jets with a tail of two-pronged jets naturally emerges. A peak of one-prong jets with a tail of two-pronged jets naturally emerges. A natural consequence of the QCD splitting function. The rate for a quark to emit a gluon of energy E at angle θ:

P(E, θ) dE dθ = 8αs(E) dE dθ / 3π E θ

With infrared and collinear divergences. With infrared and collinear divergences. Jets with balanced prongs are above. Jets with asymmetric prongs are below.

The Fractal Dimension of Jets

The correlation (fractal) dimension of the dataset is defined with pairwise distances:

dim(Q) = \lim_{\epsilon \to 0} \frac{\ln \left( \sum_{(i,j)} \text{EMD}(Q_i, Q_j) < \epsilon \right) \ln N}{\ln \epsilon}

The correlation (fractal) dimension of the dataset is defined with pairwise distances:

dim(Q) = \lim_{\epsilon \to 0} \frac{\ln \left( \sum_{(i,j)} \text{EMD}(Q_i, Q_j) < \epsilon \right) \ln N}{\ln \epsilon}

Jets become more complex at lower energies. Jets are "more than fractal" since the correlation dimension doesn't level off. We can begin to theoretically calculate it!

Towards Anomaly Detection

The lack of new physics at the LHC has stimulated interest in model-independent anomaly detection. Using the metric, we can identify the "most typical" and "least typical" jets based on their average distance to the dataset.

Q̃(E) = \frac{1}{R} \sum_{i=1}^{N} \text{EMD}(E, E_i)

A step towards anomaly detection at the LHC.

Selected References


Contact Information
Eric M. Metodiev
email: eric.metodiev@gmail.com
web: ericmetodiev.com

Visualizing Substructure

The substructure of jets is traditionally probed by computing histograms of "observables".

The most representative jets in each bin, determined via the metric, illustrate the physics that governs the observable.

The Jet Mass m probes how "wide" the jet is.

Jets with balanced prongs are above. Jets with asymmetric prongs are below.