Cell Reweighting Algorithms for Pathological Weight Mitigation Using Optimal Transport

Rishabh Jain[1] Matt LeBlanc[1,2] Jennifer Roloff[1]

[1]Department of Physics, Brown University, Providence, RI, USA
[2]The NSF AI Institute for Artificial Intelligence and Fundamental Interactions (IAIFI), Cambridge, MA, USA

The Problem of Negative Weights

- When generating next to leading order (NLO) and beyond samples, many generators produce **negatively weighted events** due to the numerical techniques used in precision calculations
  - Increases the number of samples needed for statistical precision
  - Pathological in many machine learning applications since loss functions (such as Binary Cross Entropy) expect positively weighted data

For example, consider a dataset where a fraction f_n of events are negatively weighted. The number of events $N(f_n)$ needed to reach a certain statistical significance with respect to a dataset with all positive events $N(0)$ is [1]:

$$\frac{N(f_n)}{N(f_n=0)} = \frac{1}{(1-2f_n)^2} \tag{1}$$

The computational load quickly blows up as the negative fraction approaches $\frac{1}{2}$.

The Cell Resampling Algorithm

The goal is to reweight events without affecting physical observables. The cell resampling method first introduced in [1]. We give a summary here:

1. First, find a negative event. This is the 'seed' event.
2. Build an extremely small sphere in event space around the event so only the seed is inside the sphere. This is known as the 'cell'. Now gradually extend the radius of the cell so more events are inside until eventually the total weight inside is positive.
3. Perform the following transformation on all weights inside the cell: $w_i \rightarrow \frac{\sum_{j\in C} w_j}{\sum_{j\in C}|w_j|}|w_i|$

The sum of weights is conserved inside cells. When there are many events, the cell size needed for reweighting falls below what is resolvable experimentally, **preserving physical observables** [1].

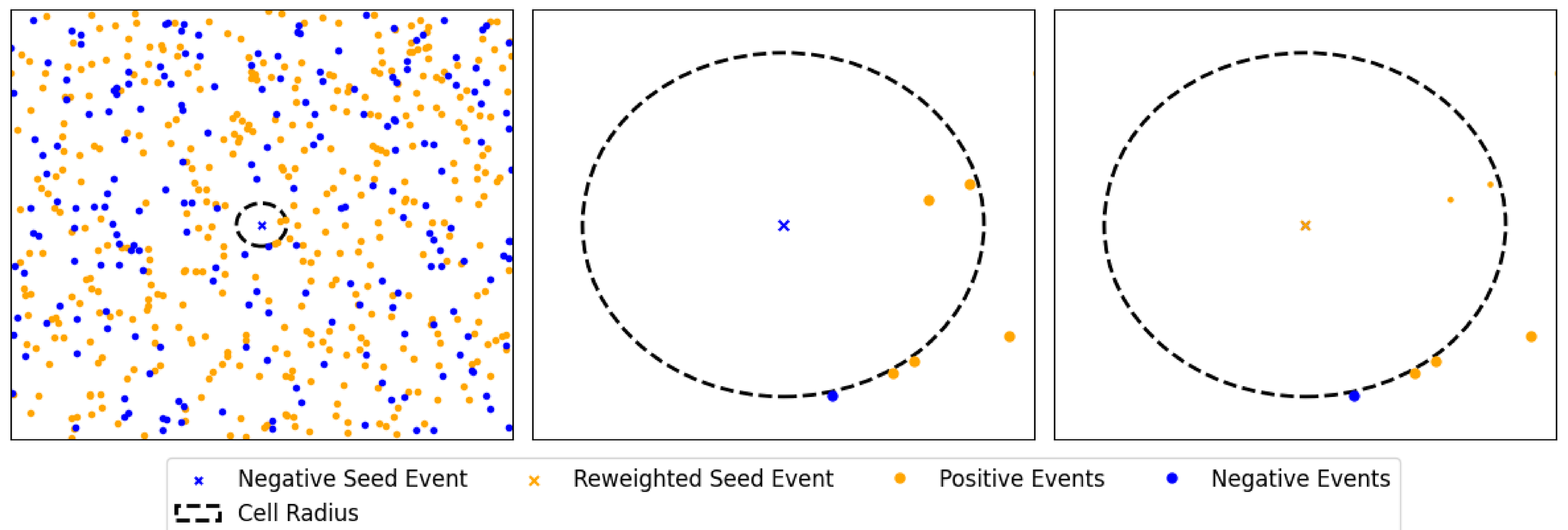


Figure 1. First pick out a negative event to be the 'seed' (left). Construct a cell around it so that the sum of weights inside the cell is positive (middle). Apply transformation on all weights inside cell (right). Now all events in the cell are positively weighted

This methodology is **process and analysis agnostic**. It is applied *post hoc*, after the events are generated [1]. We work with 10^4 events of $pp \rightarrow Zjj$ generated at NLO with Madgraph and showered and hadronized with Pythia.

Being Precise with Distance: The EMD

We choose the Energy Mover's Distance (EMD) as the metric between collider events. The EMD takes the following form [2]:

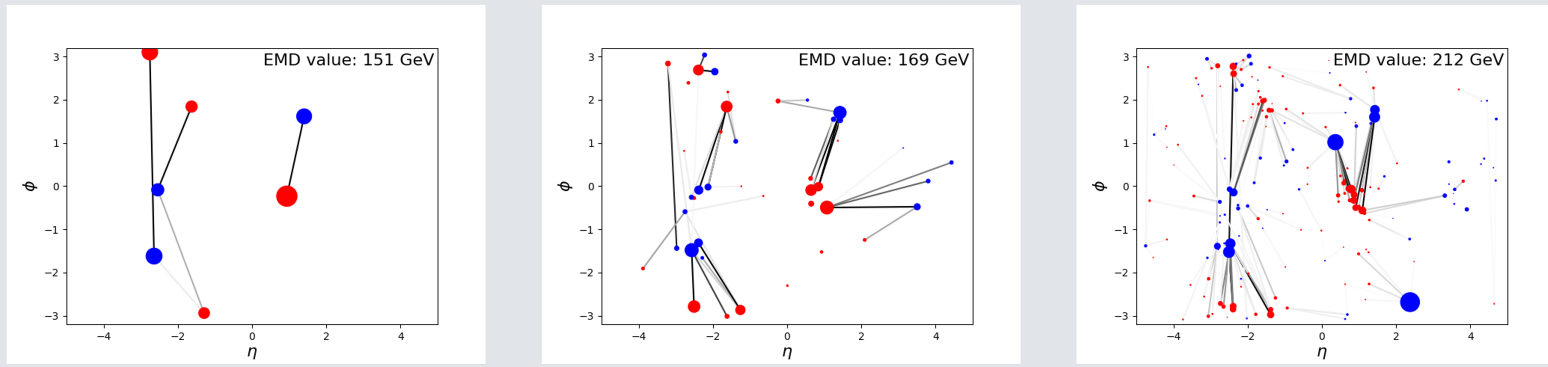
$$\mathrm{EMD}(\mathcal{E},\mathcal{E}') = \min \sum_{i=1}^{M}\sum_{j=1}^{M'} f_{ij}\left(\frac{\theta_{ij}}{R}\right)^{\beta} + \left|\sum_{i=1}^{M} E_i - \sum_{j=1}^{M'} E'_j\right| \tag{2}$$

where $\mathcal{E}, \mathcal{E}'$ are the events, f_{ij} is the transport plan, θ_{ij} is the ground metric, R determines the impact of creating/destroying energy for unbalanced OT, and β is the angular exponent. The $M(M')$ particles in event $\mathcal{E}(\mathcal{E}')$are indexed by $i(j)$ and have energies of $E_i(E'_j)$.

There are several reasons to choose the EMD as the metric between events [2]:

- EMD is **infrared and collinear (IRC) safe** $\rightarrow$ reweighting can be done at the particle level
  - Previous implementations of the reweighting scheme used IRC unsafe metrics, requiring jet clustering before cell resampling [1]
- EMD lifts the ground metric **faithfully** $\rightarrow$ no distortions between shapes.
- EMD follows all properties of a metric

We have used the parameters $\beta = 1$ and $R = 1$ with a Euclidean ground metric in the $\eta\phi$ plane.



(a) Born Level Events (b) Showered Events (c) Hadronized Events

Figure 2. Transport plan and EMD value between two events at Born level, after showering, and after hadronization.

Effect of Reweighting on Observables

To test the reweighting procedure on observables we compare reweighted samples against a reference sample with 10^7 events. Two observables are shown, h_T and $\Delta R_{Z,j_1}$.

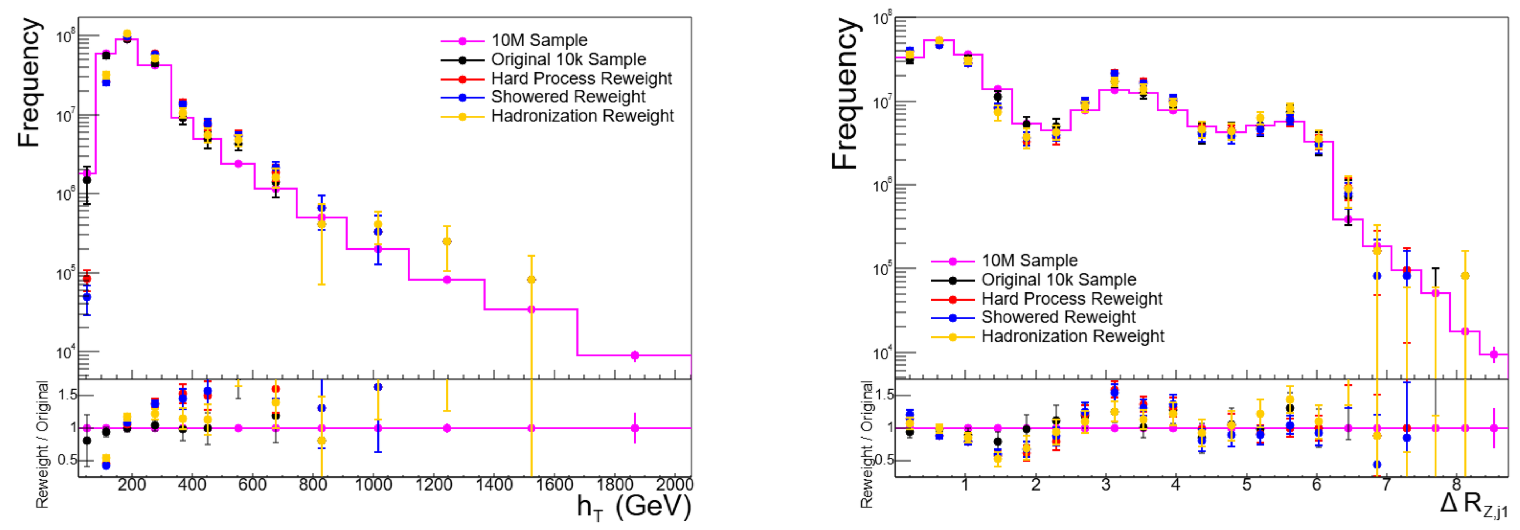
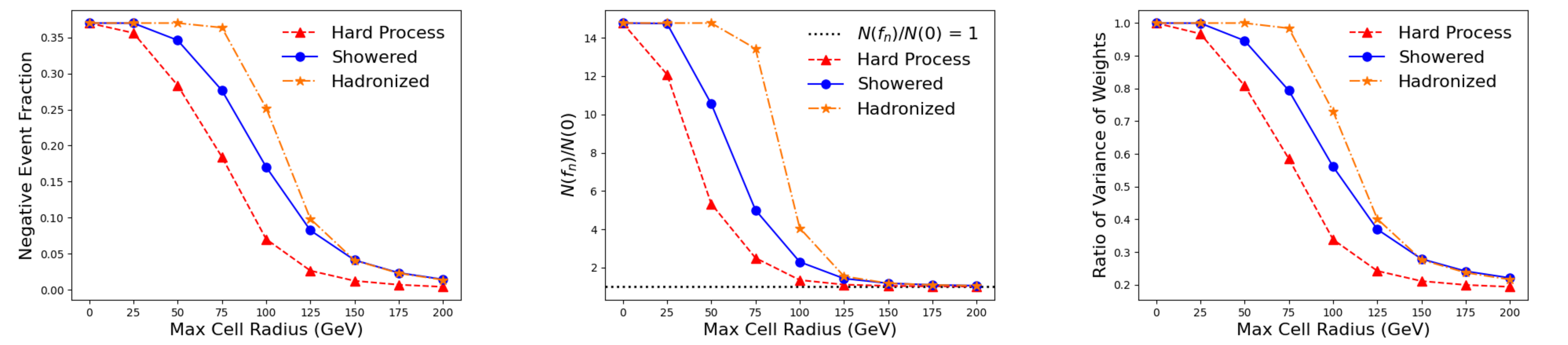


Figure 3. h_T (left) and $\Delta R_{Z,j_1}$ (right) with reference and reweighted samples with 100 GeV cell radius.

- Some bias with reweighted samples, but are **in agreement** with original sample (with uncertainties) for most bins
- Bias will decrease with larger sample size

How Effective is this Really?

The original sample has ∼37% negatively weighted events. This means ∼**14 times** more data is needed to reach the same statistical precision with respect to a dataset with all positive events.



(a) Negative event fraction (b) Statistical Significance of Sample (c) Variance of Weights

Figure 4

- As the cell radius increases, more negative events are reweighted (Figure 4 left)
- Statistical significance of the sample **improves** (Figure 4 middle) until it asymptotes to $N(f_n)/N(0) = 1$.
- The variance of the weights (Figure 4 right) **decreases**, suggesting that the weights are more clustered around the mean, rather than being very spread apart.
- However, increasing cell radius **increases bias** in observables

Conclusion and Future Steps

Applying cell resampling with OT reduces the amount of negative weights and improves statistical power of the sample. The size of the bias in observables is an effect of the small sample size.

The application of OT to cell resampling is very rich and there are several future directions with this that we are currently pursuing.

- Applying reweighting scheme on larger dataset
- Using the **Spectral** Energy Mover's Distance (SEMD) [3].
- Changing the values of R and β in the EMD.
- Using machine learning to predict EMD values for faster computation.
- Changing the ground metric from Euclidean to any of the other L_p norms.

Acknowledgements

We would like to thank the Brown University Data Science Institute for supporting this work with the seed grant 'Staying Positive: Eliminating Negative Sample Weights Using Optimal Transport' during 2024-2025.

References

[1] J. R. Andersen and A. Maier. Unbiased elimination of negative weights in Monte Carlo samples. *The European Physical Journal C*, 82(5):433, May 2022.

[2] P. T. Komiske, E. M. Metodiev, and J. Thaler. The Metric Space of Collider Events. *Physical Review Letters*, 123(4), July 2019. arXiv:1902.02346 [hep-ph].

[3] A. J. Larkoski and J. Thaler. A Spectral Metric for Collider Geometry, Aug. 2023. arXiv:2305.03751 [hep-ph].

rishabh_jain@brown.edu