# Finding NEEMo: Geometric Fitting using Neural Estimation of the Energy Mover's Distance

**Ouail Kitouni, Niklas Nolte, Mike Williams**
The NSF AI Institute for Artificial Intelligence and Fundamental Interactions
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{kitouni,nnolte,mwill}@mit.edu

## Abstract

We present a new and interesting application for networks that enforce a strict upper bound on their Lipschitz constant: geometrical fitting through differentiable estimation of the Earth Mover's Distance. We focus specifically on the field of high-energy physics, where it has been shown that a metric for the space of particle-collider events can be defined with the Earth Mover's Distance, referred to in this context as Energy Mover's Distance (EMD). This metrization has the potential to revolutionize data-driven collider phenomenology. The work presented here represents a major step towards realizing this goal by providing a differentiable way of directly calculating the EMD. We show how the flexibility that our approach enables can be used to develop novel clustering algorithms.

## 1 Introduction

The Earth Mover's Distance, otherwise referred to as Wasserstein-1 distance, is a metric defined between two probability measures. In the field of high-energy particle physics, a modified version of the Earth Mover's distance, the Energy Mover's Distance (EMD), serves as a metric for the space of collider events by defining the *work* required to rearrange the radiation pattern of one event into another Komiske et al. [2019]. In particular, the EMD is intimately connected to the structure of *infrared- and collinear-safe* observables used in the ubiquitous task of clustering particles into *jets* Komiske et al. [2020], and is foundational in the SHAPER tool for developing geometric collider observables Gambhir et al. [2022].

Recently, a novel neural architecture was developed that enforces an exact upper bound on the Lipschitz constant of the model by constraining the norm of its weights in a minimal way, resulting in higher expressiveness than other methods Kitouni et al. [2021], Anil et al. [2019]. Here, we employ this architecture—leveraging its improved expressiveness for 1-Lipschitz continuous networks—to replace the $\epsilon$-Sinkhorn estimation of the EMD in SHAPER Feydy et al. [2018], Gambhir et al. [2022] by directly calculating the EMD using the Kantorovic-Rubenstein (KR) dual formulation. The KR duality casts the optimal transport problem as an optimization over the space of 1-Lipschitz functions, which we parameterize with dense neural networks using the architecture from Kitouni et al. [2021]. With small modifications to the KR dual formulation, we are able to reliably and accurately obtain the EMD and Kantorovic potential in a differentiable way, without any $\epsilon$ approximations. This makes it possible to run gradient-based optimization procedures over the exact EMD (see Fig. 1). In addition, we expect these improvements could potentially have a major impact on jet studies at the future Electron-Ion Collider, where traditional clustering methods are not optimal Arratia et al. [2021], and more broadly in optimal transport problems.

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

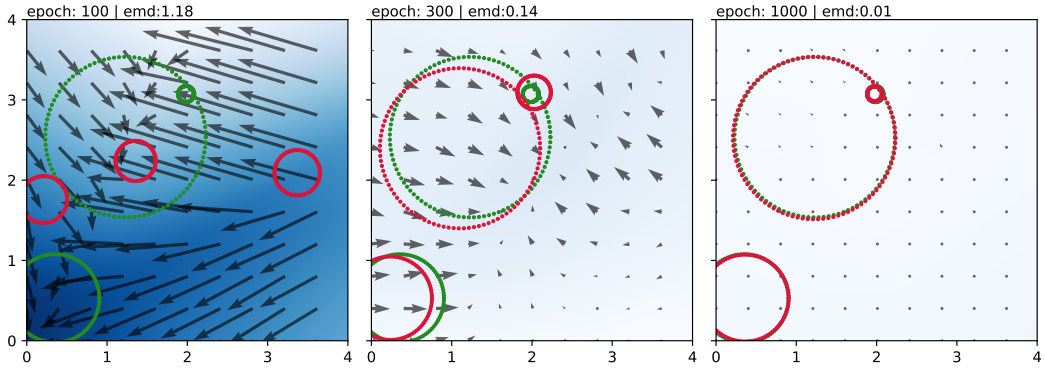


Figure 1: Fitting three synthetic clusters (green) with three circles (red) using NEEMo (see Sec. 3). The heatmap is the Kantorovic potential, parameterized as a Lipschitz-bounded network, which induces forces on the circles (shown as arrows) that drive them into perfect alignment with the target distribution (only a few steps in the evolution of the fit are shown).

## 2 Lipschitz Networks and the Energy Mover's Distance

**Lipschitz Networks** Fully connected networks can be Lipschitz bounded by constraining the matrix norm of all weights Kitouni et al. [2021], Gouk et al. [2020], Miyato et al. [2018]. Constraints with respect to a particular $L^p$ norm will be denoted as Lip$^p$. We start with a model $f(\boldsymbol{x})$ that is Lip$^p$ with Lipschitz constant $\lambda$ *i.e.*, $\forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$:

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq \lambda \|\boldsymbol{x} - \boldsymbol{y}\|_p \,. \tag{1}$$

Without loss of generality, we take $\lambda = 1$ (rescaling the inputs would be equivalent to changing $\lambda$). We recursively define the layer $l$ of the fully connected network of depth $D$ with activation $\sigma$ as

$$\boldsymbol{z}^l = \boldsymbol{W}^l \sigma(\boldsymbol{z}^{l-1}) + \boldsymbol{b}^l, \tag{2}$$

where $\boldsymbol{z}^0 = \boldsymbol{x}$ is the input and $f(\boldsymbol{x}) = z^D$ is the output of the neural network. We have that $f(\boldsymbol{x})$ satisfies equation 1 if

$$\|\boldsymbol{W}^i\|_\infty \leq 1 \quad \text{when} \quad 2 \leq i \leq D \quad \text{and} \quad ||\boldsymbol{W}^1||_{p,\infty} \leq 1 \tag{3}$$

and $\sigma$ has a Lipschitz constant less than or equal to 1. Here, $||\boldsymbol{W}||_{p,q}$ denotes the operator norm with norm $L^p$ in the domain and $L^q$ in the co-domain. It is shown in Anil et al. [2019] that when using the **GroupSort** activation, $f(\boldsymbol{x})$ can approximate any Lip$^p$ function arbitrarily well, making weight-normed networks universal approximators. An implementation of the weight constraint along with a number of examples is provided in https://github.com/niklasnolte/MonotOneNorm.

**Energy Mover's Distance** The EMD is a metric between probability measures $\mathbb{P}$ and $\mathbb{Q}$. Using the standard Wasserstein-metric notation, the EMD is defined as

$$\text{EMD}(\mathbb{P}, \mathbb{Q}) = \inf_{\gamma \in \Pi(\mathbb{P}, \mathbb{Q})} \mathbb{E}_{(x,y)\sim\gamma}\big[||x - y||_2\big], \tag{4}$$

where $\Pi(\mathbb{P}, \mathbb{Q})$ is the set of all joint probability measures whose marginals are $\mathbb{P}$ and $\mathbb{Q}$. The EMD optimization problem can be cast as an optimization over Lipschitz continuous functions using the Kantorovich-Rubinstein duality:

$$\text{EMD}(\mathbb{P}, \mathbb{Q}) = \sup_{||f||_L \leq 1} \mathbb{E}_{x\sim\mathbb{P}}\big[f(x)\big] - \mathbb{E}_{x\sim\mathbb{Q}}\big[f(x)\big], \tag{5}$$

where $f$ is Lip$^2$ continuous, *i.e.,* $||\nabla f||_2 \leq 1$. In high-energy particle collisions, the EMD is defined by using the energies of individual particles in place of probabilities, with their momentum directional coordinates representing the supports of the probability distribution. For more details, including on how unequal total energies are handled, see Komiske et al. [2019]. By performing optimizations over a constrained set of $\mathbb{P}$s, one can use the EMD to define observables over $\mathbb{Q}$.

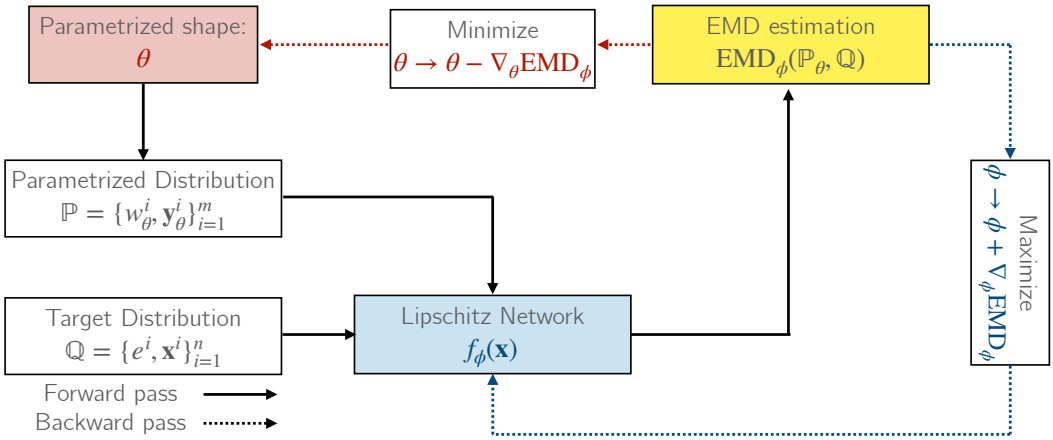


Figure 2: Training procedure to fit a parameterized shape $\mathbb{P}_{\boldsymbol{\theta}}$ to a distribution $\mathbb{Q}$. NEEMo replaces the $\epsilon$-Sinkhorn estimation in the standard SHAPER procedure with a Lipschitz network that evaluates the Kantorovic potential to obtain the EMD.

## 3 NEEMo: Neural Estimation of the Energy Mover's Distance

**Algorithm** Consider a high-energy particle-collision event with $n$ particles. Let $E^i$ be the energy of particle $i$, $\boldsymbol{x}^i$ be the direction of its momentum, and $\mathbb{Q} = \{(E^i, \boldsymbol{x}^i)\}_{i=1}^n$ be the set of all particles in the event. Following the SHAPER prescription Gambhir et al. [2022] for defining an observable $O(\mathbb{Q})$, we first define $\mathbb{P}_{\boldsymbol{\theta}} = \{w_{\boldsymbol{\theta}}^i, \boldsymbol{y}_{\boldsymbol{\theta}}^i\}_{i=1}^m$ to be any collection of points parameterized by $\boldsymbol{\theta}$, *e.g.*, these points can be sampled from any geometric object with any density distribution. The EMD between the event $\mathbb{Q}$ and the geometric object $\mathbb{P}_{\boldsymbol{\theta}}$ can be computed with equation 5 as

$$\text{EMD}(\mathbb{P}_{\boldsymbol{\theta}}, \mathbb{Q}) = \max_{\boldsymbol{\phi}} \left[ \sum_{i=1}^{n} E^i f_{\boldsymbol{\phi}}(\boldsymbol{x}^i) - \sum_{i=1}^{m} w_{\boldsymbol{\theta}}^i f_{\boldsymbol{\phi}}(\boldsymbol{y}_{\boldsymbol{\theta}}^i) \right], \tag{6}$$

where $f_{\boldsymbol{\phi}}(x)$ is a 1-Lipschitz neural network with parameters $\boldsymbol{\phi}$. At $\boldsymbol{\phi}^*$ the expression above is maximized and $f_{\boldsymbol{\phi}^*}$ is the Kantorovic potential from which the EMD is obtained as the RHS of equation 6. Since $f$ is differentiable, the optimum can be obtained using standard gradient descent techniques. This is the key improvement of NEEMo over SHAPER, which can only estimate the Kantorovic potential and the EMD up to a specified order $\epsilon$. Note that in equation 6 the expectation is computed exactly but optimization can also be done stochastically by sampling from the discrete distributions with probabilities $\{E^i\}_i$ and $\{w_\theta^i\}_i$ and using the empirical mean to estimate the EMD. This can improve convergence in some cases.

Given that all of our operations are differentiable, gradients can flow back to $\mathbb{P}_{\boldsymbol{\theta}}$. Therefore, one can also optimize the parameters $\boldsymbol{\theta}$ to obtain the best-fitting collection of points in that class. We obtain the following minimax optimization problem:

$$O(\mathbb{Q}) = \min_{\boldsymbol{\theta}} \max_{\boldsymbol{\phi}} \left[ \sum_{i=1}^{n} E^i f_{\boldsymbol{\phi}}(\boldsymbol{x}^i) - \sum_{i=1}^{m} w_{\boldsymbol{\theta}}^i f_{\boldsymbol{\phi}}(\boldsymbol{y}_{\boldsymbol{\theta}}^i) \right], \tag{7}$$

where $O(\mathbb{Q})$ quantifies how well the event $\mathbb{Q}$ is described by the class of geometric object $\mathbb{P}$ Komiske et al. [2020], Gambhir et al. [2022].

**Limitations** Unlike the conventional clustering algorithms used in high-energy particle physics, NEEMo relies on nonconvex gradient-based optimization of a neural network and a set of geometric parameters. This results in the clustering procedure itself being relatively slow and not easily implemented in real time. This problem can be alleviated with powerful custom optimizers and initialization techniques to guarantee fast convergence, though whether NEEMo could ever be run online during data taking is an open question. We note that for many potential applications, *e.g.* at the Electron-Ion Collider, this is not a problem since running online is not required.

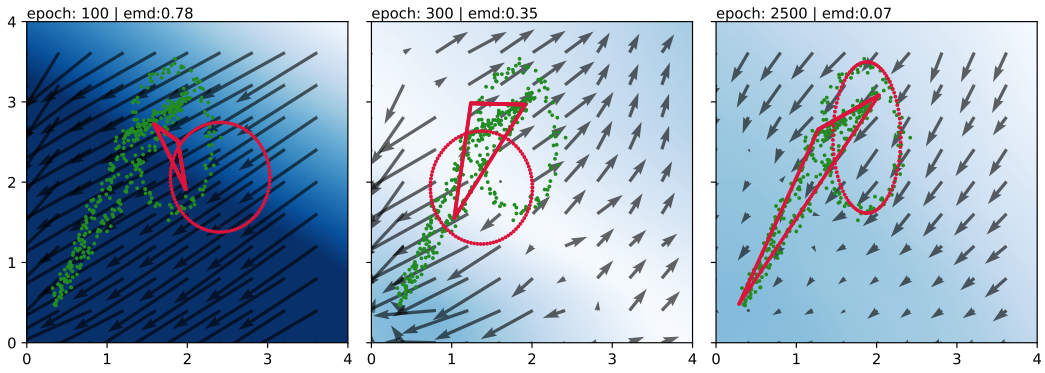


Figure 3: Same as Fig. 1, but fitting to distributions parameterized by a triangle and an ellipse.

## 4 Experiments

**Synthetic Data** We start with a few toy examples. First, consider an event consisting of three sets of particles distributed uniformly along the perimeters of circles. Here, we know the exact parameterization of our target distribution and use NEEMo to fit three randomly initialized circles to the event. Figure 1 shows a few steps in the fit evolution. The Kantorovic potential given by the Lipschitz-constrained network induces forces on the parameters of $\mathbb{P}$, which drive it to evolve from its random initialization to perfect alignment with the target distribution. In this example, $O(\mathbb{Q})$ in equation 7 quantifies the *3-circliness* of the event $\mathbb{Q}$, an observable first defined in Gambhir et al. [2022]. To highlight the flexibility, we next consider an event with two sets of particles distributed along the perimeters of a triangle and ellipse, respectively. Figure 3 shows that $\mathbb{P}$ again evolves following the gradients of the Kantorovic potential to perfect alignment with the target distribution.

**N-Subjets** We now perform a model jet-substructure study, clustering synthetic data into $N$*-subjets*. First, we generate jets with 3, 4, or 5 subjet centers distributed uniformly. From each center we generate 10 particles drawn from a Gaussian distribution. We then use our algorithm to fit 3, 4, or 5 centers to the simulated jets. Figure 4 shows that our algorithm is able to estimate the correct number of subjets. The EMD of the N-subjet fit is clearly lowest for jets with N true clusters.

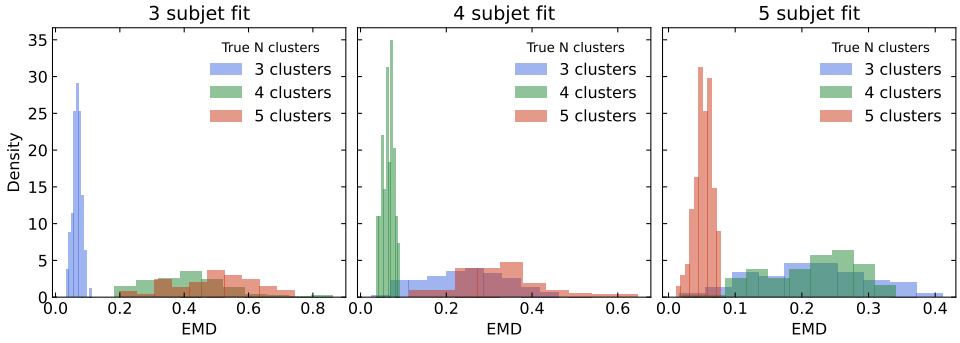


Figure 4: From left to right: Fit of N subjets (centers) to jets with 3, 4, or 5 number true subjets.

**Future Directions** In the framework developed in these proceedings, any parameterized source distribution can be chosen to fit any target distribution using the EMD, without any $\epsilon$-approximations. This can be used, *e.g.*, for constructing precision jet observables that are sensitive to percent-level fluctuations for new physics searches at LHC experiments. In addition, NEEMo provides a more precise way to quantify event modifications due to hadronization and detector effects. Finally, the flexibility provided by NEEMo could potentially have a major impact on jet studies at the future Electron-Ion Collider, where traditional clustering methods are not optimal. Rather than modifying the metric used in a sequential-recombination algorithm as in Arratia et al. [2021], the jet geometry itself can be altered using NEEMo in an event-by-event unsupervised manner. We plan to report on all of these novel directions in a follow-up journal article that is currently in preparation.

## 5 Broader Impacts

Comparing probability distributions is a fundamental task in statistics. Most commonly used methods only compare densities in a point-wise manner, whereas the Earth Mover's Distance accounts for the geometry of the underlying space. This is easily visualized in our figures showing the Kantorovic potential. Due to space constraints we only showed a few toy example applications in collider physics, but we stress that the approach we present here—directly calculating the EMD using the Kantorovic-Rubenstein dual formulation—can be applied to any optimal transport problem. While the existence of the KR duality has long been known, it only recently became possible to simultaneously enforce the *exact* 1-Lipschitz bound while achieving enough expressiveness to find the optimal Kantorovic potential. Our approach now makes it possible to perform gradient-based optimizations over the exact Earth Mover's Distance. Given the sizable impact of similar approximate methods, we expect our exact approach could have applications across many fields and types of problems.


## Acknowledgements

We thank Rikab Gambhir and Jesse Thaler for helpful discussions about SHAPER, and for introducing us to this problem. This work was supported by NSF grant PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/) and by DOE grant DE-FG02-94ER40818.


## References


Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. Metric space of collider events. *Physical Review Letters*, 123(4), jul 2019. doi: 10.1103/physrevlett.123.041801. URL `https://doi.org/10.1103%2Fphysrevlett.123.041801`.

Patrick T. Komiske, Eric M. Metodiev, and Jesse Thaler. The Hidden Geometry of Particle Collisions. *JHEP*, 07:006, 2020. doi: 10.1007/JHEP07(2020)006.

Rikab Gambhir, Akshunna Dogra, Jesse Thaler, Demba Ba, and Abiy Tasissa. Can you hear the shape of a jet? 14th International Workshop on Boosted Object Phenomenology, Reconstruction, Measurements, and Searches in HEP, 2022. URL `https://indi.to/rbQ5j`.

Ouail Kitouni, Niklas Nolte, and Mike Williams. Robust and provably monotonic networks. In *35th Conference on Neural Information Processing Systems, the Workshop on Machine Learning for the Physical Sciences*, 2021. URL `https://arxiv.org/abs/2112.00038`.

Cem Anil, James Lucas, and Roger Grosse. Sorting out Lipschitz function approximation. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 291–301. PMLR, 09–15 Jun 2019. URL `http://proceedings.mlr.press/v97/anil19a.html`.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences, 2018. URL `https://arxiv.org/abs/1810.08278`.

Miguel Arratia, Yiannis Makris, Duff Neill, Felix Ringer, and Nobuo Sato. Asymmetric jet clustering in deep-inelastic scattering. *Phys. Rev. D*, 104(3):034005, 2021. doi: 10.1103/PhysRevD.104.034005.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J. Cree. Regularisation of neural networks by enforcing lipschitz continuity, 2020.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral Normalization for Generative Adversarial Networks. *arXiv e-prints*, art. arXiv:1802.05957, February 2018.

## Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code in supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Experimental details are in the code provided.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] In the subjet experiments, the histograms are over 100 seeds.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] The fits shown along with the plots can be generated using the code provided on a laptop CPU in a few minutes.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] Standard python libraries and MIT Licensed software.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Supplemental material contains code for the main algorithm.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]