Linearized optimal transport for collider events

Tianji Cai, Junyi Cheng, and Nathaniel Craig
Department of Physics, University of California, Santa Barbara, California 93106, USA

Katy Craig
Department of Mathematics, University of California, Santa Barbara, California 93106, USA

(Received 6 October 2020; accepted 30 November 2020; published 29 December 2020)

We introduce an efficient framework for computing the distance between collider events using the tools of Linearized Optimal Transport (LOT). This preserves many of the advantages of the recently introduced Energy Mover's Distance, which quantifies the work required to rearrange one event into another, while significantly reducing the computational cost. It also furnishes a Euclidean embedding amenable to simple machine learning algorithms and visualization techniques, which we demonstrate in a variety of jet tagging examples. The LOT approximation lowers the threshold for diverse applications of the theory of optimal transport to collider physics.

DOI: 10.1103/PhysRevD.102.116019

I. INTRODUCTION

What is the distance between collider events? This question, although simple to pose, is notoriously difficult to answer. Identical events at parton level can appear to differ upon reconstruction due to soft or collinear emission, while topologically distinct events at parton level can appear identical upon reconstruction, depending on the degree of coarse graining. Despite such challenges, the value of a well-defined distance is clear: the comparison of collider events, or the reconstructed objects contained therein, is an essential step in extracting physics from collider data.

Significant progress was made toward defining a useful metric on the space of collider events in Ref. [1], where the "Energy Mover's Distance" (EMD) was introduced to compare the energy flow between events. Properly speaking, the Energy Mover's Distance is an adaptation of the Earth Mover's Distance, itself an example of the p-Wasserstein distance appearing in the theory of optimal transport. Intuitively, the p-Wasserstein distance between two normalized energy distributions represents the minimal amount of work required to rearrange one distribution to look like the other and may be modified (as in the EMD of Ref. [1]) to accommodate events with different total energies.

As observed in Ref. [1] and further developed in Ref. [2], the EMD has numerous applications to collider physics. Among other things, it provides a new perspective on existing jet variables, implies inequalities satisfied nonperturbatively by jet observables, and enables the definition of a distance between theories (where theories are defined as collections of events weighted by cross sections). From a practical perspective, the EMD defines new quantities associated with collider events that can be used as input to machine learning (ML) algorithms and leveraged in collider analyses, providing a novel intermediary between simple analytic variables and deep neural networks. The EMD defined in Ref. [1] has been subsequently applied to distance-based analysis of jets in CMS Open Data [3], to the definition of a new "event isotropy" shape variable [4], as a metric for variational autoencoder-based anomalous jet tagging [5], and (with suitable generalization) to discrimination at the full event level [6]. A number of other metrics for collider events have been explored in Ref. [7]. Broadly speaking, the many applications of the EMD pursued in Refs. [1–6] highlight the potential relevance of tools from the theory of optimal transport for collider physics.

However, one of the major practical challenges to the use of EMD in analyzing collider events is the computational cost; for a dataset containing N_{evt} events, computing the pairwise distance between all events is $\mathcal{O}(N_{\text{evt}}^2)$.[1] This poses a challenge given that computing the p-Wasserstein distance between two events itself takes fractions of a second, putting the calculation of EMDs between events in typical collider datasets beyond the reach of desktop computers. It is also unsuitable for use with ML methods

[1]The possibility of reducing such classical $\mathcal{O}(N_{\text{evt}}^2)$ strategies to $\mathcal{O}(N_{\text{evt}})$ quantum algorithms was pointed out in Ref. [8].

that require more structure than just the pairwise distances between events.

In this paper, we define an efficient framework for computing the distance between collider events by applying the tools of Linearized Optimal Transport (LOT), preserving the many advantages of the EMD while significantly reducing the computational cost and furnishing a Euclidean embedding suitable for use in a wide range of ML algorithms. In particular, we implement the LOT approximation of the 2-Wasserstein distance, as introduced in Ref. [9]. To the extent that the 2-Wasserstein distance has a pseudo-Riemannian structure (unlike p-Wasserstein distances with $p \neq 2$, including the $p = 1$ Earth Mover's Distance), the LOT approximation amounts to projecting onto the 2-Wasserstein tangent plane at a chosen reference event and computing simpler ℓ^2 distances on that plane. We make this point of view rigorous in the Appendix, where we prove that, as the reference event in the LOT approximation is refined, LOT converges to the distance between events on the tangent plane, which provides a well-defined metric on the space of events.

The LOT approach vastly speeds up the computation of optimal transport distances between collections of N_{evt} events by requiring the determination of only $\mathcal{O}(N_{\text{evt}})$ computationally intensive p-Wasserstein distances, followed by $\mathcal{O}(N_{\text{evt}}^2)$ computationally efficient ℓ^2 distances.[2] In practice, replacing the traditional optimal transport computation with this linear version reduces the computational effort of the classification task from a computer cluster to a single PC. Even with this dramatic reduction in computational time, we still achieve comparable accuracy to previous work using the original Wasserstein distances on the classification task.

Beyond the significant computational speedup, LOT provides an isometric linear embedding into Euclidean space, suitable for use in a wider range of ML algorithms. We demonstrate its utility as input to ML algorithms tasked with discriminating between samples of boosted jets containing diverse Standard Model (SM) and beyond-Standard Model (BSM) particles. Due to the fact that our ML models lack the expressivity of deep neural networks, they will not, in general, achieve the same levels of accuracy. Instead, our approach offers a much clearer interpretation in terms of the underlying physics, while still achieving very good levels of accuracy. For example, it can provide answers to questions regarding what properties are most important in distinguishing them from each other; see Fig. 5.

This paper is organized as follows. In Sec. II, we review the p-Wasserstein distance and the Linearized Optimal Transport approximation to the $p = 2$ distance, framed in terms suitable for application to collider events. We then illustrate features of the LOT approximation in the context of jet tagging in Sec. III, computing LOT pseudodistances between various classes of boosted jets using an isotropic (in cylindrical coordinates) distribution as a reference event. The utility of LOT as an input to simple machine learning algorithms is highlighted in Sec. IV, where we explore the performance of linear discriminate analysis (LDA), k-nearest neighbor (kNN), support vector machine (SVM), and k-medoids clustering algorithms in the pairwise classification of boosted QCD, W, t, Higgs, and BSM jets. The comparable performance of models respectively coupled with the LOT and EMD metrics suggests that the former approximation matches the discriminating power of the latter metric while offering considerable computational speedup. It is also readily amenable to visualization, which we demonstrate in a number of examples. We conclude and enumerate a variety of future directions in Sec. V. A proof of the convergence of the LOT approximation to a true metric in the continuum limit is reserved for the Appendix.

II. LINEARIZED OPTIMAL TRANSPORT

Let an event $\mathcal{E}$ denote a collection of particles at locations x_i in a rectangular domain Ω, with energies $E_i, \tilde{E}_j \geq 0$.[3] Given two events $\mathcal{E}, \tilde{\mathcal{E}}$ with the same total energy, $\sum_i E_i = \sum_j \tilde{E}_j$, the theory of optimal transport provides various notions of distance between the two events. In particular, for $p \geq 1$, the p-Wasserstein distance is given by

$$W_p(\mathcal{E}, \tilde{\mathcal{E}}) = \min_{g_{ij} \in \Gamma(\mathcal{E}, \tilde{\mathcal{E}})} \left(\sum\nolimits_{ij} g_{ij} \|x_i - \tilde{x}_j\|^p \right)^{1/p},$$
$$\Gamma(\mathcal{E}, \tilde{\mathcal{E}}) = \left\{ g_{ij} : g_{ij} \geq 0, \sum_j g_{ij} = E_i, \sum_i g_{ij} = \tilde{E}_j \right\}, \quad (1)$$

where $\|x_i - \tilde{x}_j\|$ denotes the angular distance on the underlying space Ω, which we will often refer to as the *ground metric*. When $p = 1$ or 2, W_p is also known as the Earth Mover's Distance or the Monge-Kantorovich distance, respectively. Up to normalizing the energies $E_i, \tilde{E}_j$ by dividing through by the total energy of each event, we may assume without loss of generality that the total energy of all events we consider equals 1.

One interpretation of the p-Wasserstein distance is that it represents the minimal amount of "effort" required to rearrange the distribution of energy in $\mathcal{E}$ to match $\tilde{\mathcal{E}}$. In this case, g_{ij} represents the amount of energy moved from particle i in event $\mathcal{E}$ to particle j in event $\tilde{\mathcal{E}}$, and $\|x_i - \tilde{x}_j\|^p$ represents the "cost" of moving energy between the two

[2] Another pseudo-Riemannian structure, reminiscent of the 2-Wasserstein metric, has also been used to reduce the computational complexity of multiparticle correlators [10].

[3] While the detector on which the collision data is recorded is a cylinder, due to the fact that we will translate jets clustered with unit radius parameter to be centered at the origin, we may neglect the periodic boundary conditions in the azimuthal angle and consider the underlying domain to be a rectangle.

locations. With this interpretation, $\Gamma(\mathcal{E},\tilde{\mathcal{E}})$ is the set of possible ways to rearrange $\mathcal{E}$ to look like $\tilde{\mathcal{E}}$, known as the set of *transportation plans*: any rearrangement g_{ij} can only move non-negative amounts of energy; the total amount of energy moved from a fixed particle i in $\mathcal{E}$ to all of the particles in $\tilde{\mathcal{E}}$ must coincide with the original energy E_i; and, symmetrically, the total amount of energy moved from all of the particles in $\mathcal{E}$ to any fixed particle j in $\tilde{\mathcal{E}}$ must coincide with E'_j. More generally, there are several methods to extend the Wasserstein distance to events $\mathcal{E}$ and $\tilde{\mathcal{E}}$ with different total energies, including the version of the Earth Mover's Distance considered in Ref. [1], which is a type of partial optimal transport distance [11–13] created by interpolating between the 1-Wasserstein distance and the total variation norm.

Over the past 20 years, optimal transport distances have emerged as important metrics for image classification tasks [14–19]. These metrics are unique in that they lift the ground metric on the underlying space to the set of probability distributions on that space. This is in contrast with more traditional metrics, such as the ℓ^2 norm. For example, in an image based approach, the ℓ^2 norm computes the distance between two events $\mathcal{E}$ and $\tilde{\mathcal{E}}$ by, first, binning the particles on a grid with N bins; second, representing the energy at each grid location by vectors $v,\tilde{v}\in\mathbb{R}^N$; and, third, computing the distance between $\mathcal{E}$ and $\tilde{\mathcal{E}}$ via the standard Euclidean norm,

$$d_{\ell^2(\mathbb{R}^N)}(\mathcal{E},\tilde{\mathcal{E}}) \coloneqq \left(\sum_{i=1}^{N}|v_i-\tilde{v}_i|^2\right)^{1/2}. \tag{2}$$

Unlike the Wasserstein metric, the ℓ^2 norm does not respect the geometry of the underlying space. For example, suppose each event consists of a single particle with energy 1, the particles are distance $\|x_1-\tilde{x}_1\|$ apart, and the grid for the ℓ^2 norm is fine enough so that the particles fall in different bins. Then,

$$W_p(\mathcal{E},\tilde{\mathcal{E}}) = \|x_1-\tilde{x}\| \quad \text{and} \quad d_{\ell^2(\mathbb{R}^n)}(\mathcal{E},\tilde{\mathcal{E}}) = \sqrt{2}.$$

While the p-Wasserstein metrics take into account the particles' locations on the underlying space, this information is neglected by the classical ℓ^2 norm. This ability to preserve spatial information provides the p-Wasserstein metrics with a natural advantage in image classification tasks.

In spite of these theoretical benefits of optimal transport metrics, wider adoption in image classification has been slowed by two obstacles: computational cost and limited choice of classification algorithms. In terms of computational efficiency, computing the p-Wasserstein distance between two events, with n particles in each event, requires $O(n^3)$ operations via Bertsekas's auction algorithm and $O(n^2\log(n))$ operations via entropic regularization and the Sinkhorn algorithm [20–24]. This is in contrast to the classical ℓ^2 norm, which is naively $\mathcal{O}(n)$, when the number of bins is chosen proportional to the number of particles, $n\sim N$. In image classification tasks, the high cost of the p-Wasserstein metrics is compounded by the fact that one needs to compute the pairwise p-Wasserstein distances between the entire collection of N_{evt} images, requiring $O(N_{\text{evt}}^2)$ computations of the distance. In the particular case of classifying jet events, the number of particles per event is relatively small, $n\approx 10^2$, and it is this latter need to compute pairwise distances between a large number of events, $N_{\text{evt}}\approx 10^5$, which is the main computational expense. Furthermore, existing work using classical optimal transport metrics must also cope with the significant computational demands of storing the matrix of pairwise distances.

The goal of the present work is to overcome the problem of high computational cost and limited choice of algorithms by using the Linearized Optimal Transport approximation of the 2-Wasserstein distance, originally introduced by Wang *et al.* [9] as a method for visualizing variation in sets of images. Let $\mathcal{R}$ denote the reference event, a collection of particles at locations y_i with energies R_i. For any event $\mathcal{E}$, let r_{ij} denote an optimal transport plan from $\mathcal{R}$ to $\mathcal{E}$, that is, a minimizer of (1). (Note that there may be more than one optimal transport plan between two given events.) In general, a transport plan r_{ij} may send energy from particle i in the reference measure to many different particles in event $\mathcal{E}$. Consider the average of these locations, weighted by how much energy is sent to each and normalized by the amount of energy starting at particle i,

$$z_i \coloneqq \frac{1}{R_i}\sum_j r_{ij}x_j. \tag{3}$$

This provides a map from an event $\mathcal{E}$ to a vector z_i in $2n$-dimensional Euclidean space, $\mathbb{R}^{2n}$, where n is the number of particles in the reference jet.

The LOT approximation of the 2-Wasserstein metric measures the distance between two events $\mathcal{E}$ and $\tilde{\mathcal{E}}$ by considering the Euclidean distances between all pairs $(z_i,\tilde{z}_i)$, weighted by the mass starting at particle i,

$$\text{LOT}_{r,\tilde{r}}(\mathcal{E},\tilde{\mathcal{E}}) = \left(\sum\nolimits_i R_i\|z_i-\tilde{z}_i\|^2\right)^{1/2}. \tag{4}$$

Note that this approximation depends on the choice of transport plans $r_{ij},\tilde{r}_{ij}$.

In Fig. 1, we illustrate the LOT-W2 computation and its relationship to the standard 2-Wasserstein metric (OT-W2). The top row shows two optimal transport plans that rearrange a uniform reference jet of 81 constituent particles (green) into two sample jets (blue and red), according to the

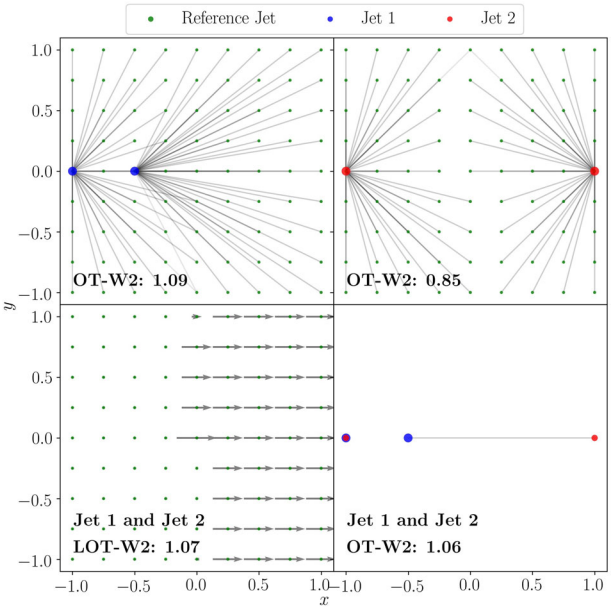

FIG. 1. Upper left: an optimal movement using the OT-W2 metric to rearrange a uniform reference jet of $9 \times 9 = 81$ constituent particles (green) into the sample jet 1 (blue). Upper right: an optimal movement using the OT-W2 metric to rearrange the same uniform reference jet (green) into another sample jet 2 (red). Lower left: an optimal movement to rearrange the sample jet 1 into the sample jet 2 using LOT-W2. Lower right: an optimal movement to rearrange the two sample jets directly using OT-W2.

exact 2-Wasserstein metric. Gray lines indicate how energy from particle y_i in the reference jet is sent to particle x_j in sample jet 1 or particle $\tilde{x}_j$ in sample jet 2. Note that, as there are multiple optimal ways to perform this rearrangement, the rearrangement is not guaranteed to be symmetric: in the top left figure, compare the fifth particle from the left on the bottom row (which splits mass between both blue particles) to the top row (which sends all mass to the right particle). In the bottom left subplot, we illustrate $\tilde{z}_i - z_i$, to visualize the difference in how the reference jet is rearranged for jet 1 and jet 2. Predictably, we observe that the main difference is energy goes further to the right in the case of jet 2. The LOT approximation of the 2-Wasserstein distance is computed by taking the sum of the lengths of the gray vectors squared, weighted by the energy of the reference measure $R_i = 1/81$, so that $LOT_{r,\tilde{r}}(\mathcal{E}, \tilde{\mathcal{E}}) \approx 1.07$. Finally, in the lower right subplot, we illustrate the OT-W2 distance between jet 1 and jet 2, which corresponds to moving half of the energy in the jet 1 a distance 1.5, so $W_2(\mathcal{E}, \tilde{\mathcal{E}}) = (1.5^2/2)^{1/2} \approx 1.06$.

The LOT approximation does not, in general, provide a metric on the space of events. For example, if the reference event $\mathcal{R}$ consists of a single particle at location y_1, then $z_1 = \sum_j x_j E_j$ is the "center of energy" of $\mathcal{E}$, and any two events $\mathcal{E}, \tilde{\mathcal{E}}$ with equal center of energy satisfy $\mathrm{LOT}_{r,\tilde{r}}(\mathcal{E}, \tilde{\mathcal{E}}) = 0$. Consequently, it is clear that a necessary condition for the LOT approximation to capture finer properties of events is that the reference event cannot be too concentrated. In fact, this condition is also sufficient. In the Appendix, we describe how the LOT approximation extends to reference events $\mathcal{R}$ given by general measures on Euclidean space. When the reference event does not concentrate on lower -dimensional sets, the LOT approximation coincides with the *transport metric with base* $\mathcal{R}$, denoted $W_{2,\mathcal{R}}$, which is a well-defined metric on the space of events, corresponding to taking the distance between two events by projecting on the 2-Wasserstein tangent plane at $\mathcal{R}$. In Corollary 1 of the Appendix, we prove that, if the reference event $\mathcal{R}^N$ is given by a collection of N^2 particles, uniformly distributed on a rectangle Ω, with equally weighted energies $R_i^N = 1/N^2$, then, as $N \to +\infty$, the LOT approximation converges to $W_{2,\mathcal{R}}$, where $\mathcal{R}$ is the probability measure uniformly distributed on Ω,

$$\lim_{N \to +\infty} \mathrm{LOT}_{r^N, \tilde{r}^N}(\mathcal{E}, \tilde{\mathcal{E}}) = W_{2,\mathcal{R}}(\mathcal{E}, \tilde{\mathcal{E}}). \tag{5}$$

For this choice of $\mathcal{R}$ and any events $\mathcal{E}, \tilde{\mathcal{E}}$ on Ω, the transport metric is bounded above and below by the original 2-Wasserstein distance [25],

$$W_2(\mathcal{E}, \tilde{\mathcal{E}}) \le W_{2,\mathcal{R}}(\mathcal{E}, \tilde{\mathcal{E}}) \le C W_2(\mathcal{E}, \tilde{\mathcal{E}})^{2/15}, \tag{6}$$

where the constant $C > 0$ depends on Ω. In this way, LOT not only converges to a well-defined transport metric $W_{2,\mathcal{R}}$, but that transport metric captures the behavior of the original 2-Wasserstein metric at large and small distances. See Refs. [26–28] for further analysis of the LOT embedding.

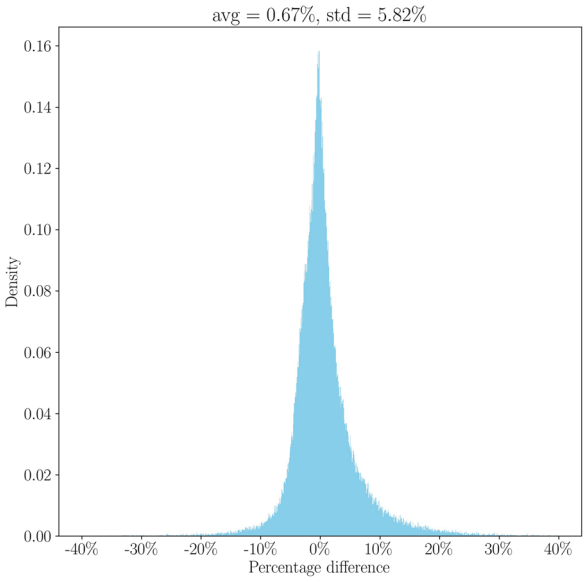

FIG. 2. Distribution of percentage differences between the LOT approximation and the 2-Wasserstein distance for pairs of events in a sample of 500 mixed W and QCD jets.

On one hand, the LOT approximation, the $W_{2,\mathcal{R}}$ transport metric, and the 2-Wasserstein metric do not need to attain similar values for the pairwise distances between events in order for LOT to offer good discrimination power in classification and clustering tasks. However, as an illustration of the similarity of LOT and the 2-Wasserstein metric in practice, we plot in Fig. 2 a histogram of the difference between the LOT approximation and the exact 2-Wasserstein metric when computing the pairwise distances between a sample of 500 mixed W and QCD jets. The reference event is given by 225 constituent particles uniformly distributed on a 15×15 grid. We observe that the LOT approximation is on average slightly larger than the 2-Wasserstein distance (mean 0.67%), and they are generally of comparable size (standard deviation equal to 5.82%).

The key benefit of the LOT approximation is that it provides a natural embedding $\mathcal{E} \mapsto z_i$ of events into

TABLE I. Results for the seven jet tagging tasks using four different machine learning models coupled with the LOT coordination.

			Comparison task						
Model	Dataset		W vs QCD	t vs QCD	t vs W	H vs QCD	H vs W	BSM vs QCD	BSM vs W
LDA	Sample dataset	**AUC**	**0.6896**	**0.7863**	**0.8464**	**0.7642**	**0.7865**	**0.7158**	**0.7244**
		TPR	0.6926	0.7746	0.7886	0.7378	0.7762	0.6713	0.6562
		FPR	0.3133	0.2020	0.0958	0.2095	0.2032	0.2397	0.2074
		Approximate run time	Several seconds						
	Full dataset	**AUC**	**0.7041**	**0.8077**	**0.8573**	**0.7703**	**0.8443**	**0.7337**	**0.7455**
		TPR	0.7156	0.7969	0.7957	0.7661	0.8254	0.7549	0.6804
		FPR	0.3075	0.1815	0.0812	0.2255	0.1368	0.2874	0.1894
		Approximate run time	Several seconds						
SVM	Sample dataset	**AUC**	**0.8410**	**0.8630**	**0.8751**	**0.8349**	**0.8831**	**0.8239**	**0.8806**
		TPR	0.8148	0.8929	0.8333	0.8006	0.8750	0.8582	0.9090
		FPR	0.1327	0.1669	0.0831	0.1308	0.1088	0.2104	0.1478
		Approximate run time	2 h						
	Full dataset	**AUC**	**0.8687**	**0.8780**	**0.8805**	**0.8426**	**0.9100**	**0.8331**	**0.9077**
		TPR	0.8451	0.8873	0.8365	0.8185	0.9103	0.8471	0.9191
		FPR	0.1077	0.1313	0.0755	0.1332	0.0904	0.1808	0.1037
		Approximate run time	6 h						
	Hyperparameters	C	1.0	1.0	10.0	1.0	1.0	1.0	1.0
		γ	100.0	100.0	10.0	100.0	100.0	100.0	100.0
kNN	Sample dataset	**AUC**	**0.8191**	**0.8450**	**0.8659**	**0.8203**	**0.8628**	**0.8026**	**0.8361**
		TPR	0.7741	0.8164	0.8040	0.7975	0.8295	0.8172	0.8241
		FPR	0.1358	0.1264	0.0723	0.1568	0.1038	0.2120	0.1520
		Approximate run time	15 min						
	Full dataset	**AUC**	**0.8455**	**0.8601**	**0.8735**	**0.8280**	**0.8831**	**0.8192**	**0.8772**
		TPR	0.8033	0.8217	0.8156	0.8040	0.8566	0.8261	0.8836
		FPR	0.1123	0.1014	0.0686	0.1479	0.0905	0.1876	0.1292
		Approximate run time	4 h						
	Hyperparameter	k	20	40	10	20	20	10	20
k-medoids clustering	Sample dataset	**AUC**	**0.6797**	**0.8096**	**0.8074**	**0.7689**	**0.8028**	**0.7622**	**0.6698**
		TPR	0.7947	0.9282	0.6583	0.8374	0.6835	0.8837	0.5216
		FPR	0.4354	0.3089	0.0436	0.2996	0.0778	0.3592	0.1821
		Signal percentage (signal, background)	(63.78%, 25.97%)	(74.70%, 9.27%)	(94.00%, 27.02%)	(73.60%, 18.81%)	(90.11%, 26.24%)	(71.05%, 15.33%)	(74.81%, 37.75%)
		Clusters' size (signal, background)	(6118, 3882)	(6159, 3841)	(3565, 6435)	(5682, 4318)	(3861, 6139)	(6211, 3789)	(3549, 6451)
		Medoids true labels (signal: 1, background: 0)	(1, 0)	(0, 0)	(1, 0)	(1, 0)	(1, 1)	(1, 0)	(1, 0)
		Approximate run time	30 min						

Euclidean space. This embedding is useful for two reasons. First, $\mathrm{LOT}_{r,\tilde{r}}(\mathcal{E},\tilde{\mathcal{E}})$ coincides with the ℓ^2 distance of the Euclidean coordinates $z_i, \tilde{z}_i$, weighted by the energies of the reference measure R_i. Consequently, to compute the pairwise LOT approximation between all events in a sample requires $O(N_{\mathrm{evt}})$ computations of the 2-Wasserstein metric, in order to construct the embedding $\mathcal{E} \mapsto z_i$, and then $O(N_{\mathrm{evt}}^2)$ computations of the ℓ^2 metric, in order to compute the value of LOT between all events. Given that each computation of ℓ^2 is, on average, 4 orders of magnitude faster than computing a Wasserstein distance, this results in an enormous computational advantage.

The second reason that the LOT Euclidean embedding is useful in jet classification is that it allows us to apply a wider range of classification algorithms directly to the vectors $z_i, \tilde{z}_i$ representing the events $\mathcal{E}, \tilde{\mathcal{E}}$. While existing work using optimal transport for jet classification considered algorithms that only rely on pairwise distances between all events, such as kNN, by using the LOT Euclidean embedding, we are able to apply algorithms that require a Euclidean structure, such as LDA. By leveraging this Euclidean structure, even this simplistic algorithm is able to provide novel ways to visualize variation in the dataset (see Fig. 5) and surprisingly accurate classification, compared to more sophisticated learning methods (see Table I). Finally, by passing the Euclidean coordinates directly to the ML models and thereby delegating computation of the entire pairwise LOT approximate distance to efficient downstream methods, the LOT approximation has a large storage advantage over traditional optimal transport techniques in ML.

III. OBJECT CLASSIFICATION WITH LOT

To demonstrate the efficacy of the LOT framework, we now focus exclusively on the task of jet tagging, that is, distinguishing one type of jet from another. In addition to being an important tool in experimental analyses, jet tagging serves as an ideal playground to test new machine learning ideas in the realm of both supervised classification and unsupervised clustering. Given that optimal transport quantifies the similarity between the energy flows of two jets, the hope is that the metrics can effectively capture the differences among a variety of jet types. For the purposes of this application, we take an event to consist of a single jet and consider the flow of p_T associated with particles in the jet.

Here, we consider five types of jets: single-pronged QCD (quark or gluon) jets, two-pronged boosted W boson jets, three-pronged boosted top quark jets, two-pronged boosted Higgs boson jets, and two-pronged boosted jets from a hypothetical new particle. This new BSM particle ϕ is taken to be a scalar transforming in the 6 representation of $SU(3)_C$ and carrying electromagnetic charge $+\frac{1}{3}$; we consider a benchmark mass of $m_\phi = 100$ GeV with a width of $\Gamma_\phi = 2$ GeV. It couples equally to all quark pairs that respect charge conservation. We calculate the Feynman rules for this BSM particle ϕ using FEYNRULES [29].

Instead of examining all possible pairwise combinations, we narrow our analysis to the following seven pairs: W vs QCD, t vs QCD, t vs W, H vs QCD, H vs W, BSM vs QCD, and BSM vs W. For the most part, these comparisons could be thought of as treating both QCD and W boson jets as backgrounds, whereas top, Higgs boson, and BSM jets are treated as signals. The W vs QCD pair is introduced as a benchmark for the performance of the other six tagging tasks, as well as for a meaningful comparison with the results obtained in Ref. [1].

We generate proton-proton collision events using MADGRAPH2.6.7 [29] at $\sqrt{s} = 14$ TeV, where the two-pronged boosted Higgs boson jets are generated via $q\bar{q} \to Z(\to \nu\bar{\nu}) + H(\to b\bar{b})$ and the BSM jets are generated through $q\bar{q} \to \phi\bar{\phi}$; all other SM jets are created via pair production. The BSM (anti)particle subsequently decays to two quarks. The matrix elements are then fed into PYTHIA8.243 [30], with hadronization and multiple particle interactions switched on using default tuning and showering parameters. No detector simulation is included. Afterward, we cluster the jets in FASTJET3.3.2 [31] using the anti-k_T algorithm with a jet radius of 1.0, where at most two jets with $p_T \in [500, 550]$ GeV and $|y| < 1.7$ are kept.

To remove any artificial difference in the energy flows of the produced jets, every jet is preprocessed by boosting and rotating to center the jet 4-momentum and vertically align the principal component of the constituent p_T flow in the rapidity-azimuth plane using the ENERGYFLOW package [1,3,10,32,33].

In order to have a unified framework for the seven comparison tasks, we work with a single choice of reference jet. The reference jet has a total p_T of 525 GeV and 225 constituent particles, each with the same amount of p_T evenly distributed on a 15×15 grid with $|y| \leq 1.7$ and $|\phi| \leq \frac{\pi}{2}$. This corresponds to an isotropic distribution on the cylinder; note that related reference distributions were explored in Ref. [4] for the purposes of defining the event isotropy variable. While raising the number of particles in the reference jet does marginally improve the approximation of LOT to $W_{2,\mathcal{R}}$, it greatly increases the computational cost and does not improve the accuracy of the classification and clustering tasks. We have also tried nonuniform reference jets, and the resulting LOT approximation does not show any material difference compared to what is obtained from the uniform reference jet. Furthermore, as we justify rigorously in the Appendix, the LOT approximation with a uniform reference jet can be seen as an approximation of $W_{2,\mathcal{R}}$, the transport metric with base $\mathcal{R}$, which approximates the original 2-Wasserstein metric at large and small distances; see Eq. (6). For this reason, we will often refer to the LOT approximation as the LOT pseudodistance in what follows.

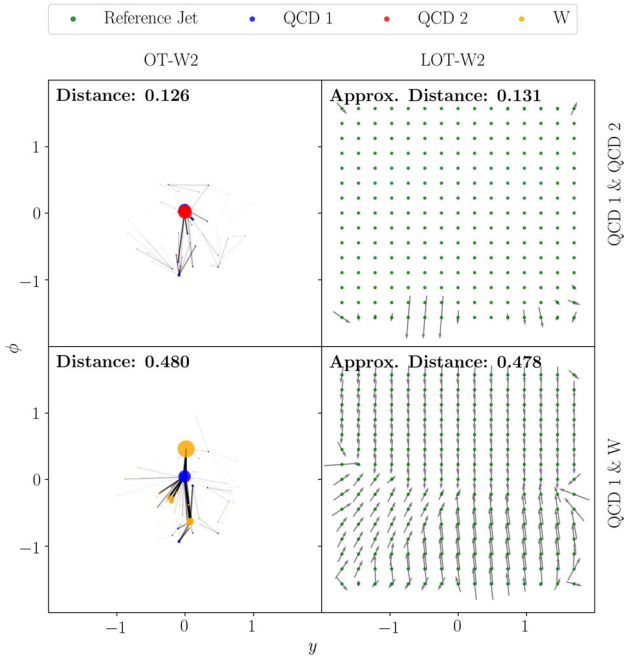

FIG. 3. Upper left: the optimal movement to rearrange one QCD jet (red) into another (blue) using the exact OT-W2 metric. Upper right: the optimal movement to rearrange the same two QCD jets using LOT-W2. Lower left: the optimal movement to rearrange a W jet (orange) into a QCD jet (blue) using the exact OT-W2 metric. Lower right: the optimal movement to rearrange the same QCD and W jets using LOT -W2.

We first normalize the p_T of all jets to unity before using the Python Optimal Transport library [34] to compute the exact Optimal Transport (OT) distance between a given jet and the reference jet, with the cost being the Euclidean distance squared in the rapidity-azimuth coordinate.[4]

Once we have this OT distance in hand, we proceed to calculate the linear embedding for each jet using the method in Sec. II. Later, we recover the approximate LOT pseudodistance between any two jets from the weighted ℓ^2 distance between their Euclidean coordinates, which we refer to as their LOT coordinates. (Note that, due to the fact that we choose our reference jet so that all particles have equal energy, the weighted ℓ^2 norm reduces to a classical ℓ^2 norm in our setting.)

Figure 3 shows the optimal energy movements between two sample QCD jets and between sample QCD and W jets using the OT-W2 distance and the LOT-W2 approximation, respectively. All jets are normalized to have unit p_T before computing both metrics. In visualizing the OT-W2 metric, points in the y-ϕ plane represent constituent particles, with sizes proportional to their p_T; the darkness of the lines connecting points in the two jets indicate how much p_T is moved from one particle to another. In visualizing the LOT pseudodistance, vectors located at each particle in the reference jet indicate the *difference* between movement of p_T from that particle in the reference jet to particles in the respective sample jets. In each case, the total distance between the two jets is also shown. These examples illustrate the qualitative properties of both metrics applied to simulated events: in the case of OT-W2, large OT distances correspond to the movement of significant amounts of energy between particles widely separated in the ground metric, while large LOT pseudodistances correspond to very different transport plans between the reference jet and the respective particles. We observe that the LOT-W2 pseudodistance is numerically close to the exact OT-W2 distance, consistent with the bounds from inequality (6).

[4]This normalization step obviates the need to modify the OT distance with an additional difference term as in Ref. [1]. For jet samples in the p_T range explored here, we found that simple machine learning algorithms exhibit comparable or slightly better performance when using exact OT-W1 or OT-W2 distances computed between normalized jets, compared to EMD distances computed between non-normalized jets.

IV. MACHINE LEARNING WITH LOT

Once we assign a LOT coordinate to each jet, the inputs for jet tagging become standardized, enabling the application of a large pool of simple machine learning algorithms. LDA, kNN, and SVM are among many suitable algorithms for classification. Such a meaningful jet representation also makes it possible to try unsupervised clustering algorithms where we leave the model itself to assign a label for each jet. One simple example is k-medoids clustering. Though relatively limited in performance, all the above-mentioned traditional models have important advantages over neural networks. They are more computationally economic, have fewer hyperparameters to tune, and offer better human interpretability. Most of them are also off-the-shelf functions implemented in the PYTHON package SCIKIT-LEARN [35], making their adoption easier in practice. In our analysis, we use all four aforementioned machine learning models to either classify or cluster the jets.

The simple supervised classifier kNN [36] relies on a majority vote of one's closest k neighbors in the training set to determine the class membership of the new data point. Here, k is a model hyperparameter to be tuned. We test k in the range from 10 to 1000 with an increment of 10. Since kNN relies only on a notion of pairwise distance, it serves as a good probe to check whether our LOT approximation sufficiently captures the difference among various jet types while at the same time adequately reflecting the similarity within one specific type. The simplicity in understanding kNN and its reliance only on pairwise distances between events contribute to its adoption in the original EMD paper [1].

A more sophisticated model, the SVM [37], lifts the inputs into a high-dimensional space and finds an optimal hyperplane to best separate the data. Key to SVM is the choice of a kernel function. Here, we use the common

radial basis function kernel $\exp[-\gamma d(x,x')^2]$, where $d(x,x')$ is the LOT pseudodistance between the two data points and γ is a tunable hyperparameter controlling how much influence a single training example has. A high γ suggests that only nearby points are considered. Another hyperparameter of the model C regulates the strength of the penalty term when a sample is misclassified, where a high value implies that nearly all training examples need to be classified correctly. In our analysis, we let both C and γ run from 10^{-5} to 10^{5} again with an increment of 10. Thus, there are $11 \times 11 = 121$ pairs of hyperparameters, and the model needs to be run for 121 times to determine the best choice.

Since both SVM and kNN involve hyperparameter tuning, they are relatively time consuming to train for large datasets. In contrast, LDA [38] has closed-form solutions with no hyperparameter, making it an attractive model for a quick first look into the data. With the assumptions that the input data is Gaussian and the Gaussian for each class shares the same covariance matrix, LDA projects the input high-dimensional data onto a direction that is most discriminative, denoted as the LDA direction. Here, we use LDA both as a classifier and as a tool for visualization, a point to be elaborated later.

For unsupervised learning, we choose as a first try k-medoids clustering [39] implemented in the PYTHON package PYCLUSTERING [40]. The goal of the model is to partition the dataset so that the distance between points labeled to be in a cluster and the point designated as the center of that cluster is minimized. Note that the centers, called medoids, are chosen from actual data points. For the present application, the model is asked to group the unlabeled data into $k = 2$ clusters. Then, the true labels are uncovered. The cluster with a higher percentage of signal jets is denoted as the signal cluster, whereas the other is designated as the background cluster. We also retrieve the true labels of the two picked medoids. Ideally, the true label of the medoid should be the same as the label of its own cluster. If not, we prefer the cluster's label. We then assign all jets in the signal cluster as signals and those in the background cluster as background jets. This assignment is compared with the ground truth to assess the performance of our clustering model. Strictly speaking, the model is semisupervised, for we need the true labels to decide which cluster is the signal cluster. A more detailed discussion of k-medoids and its performance will be given in a later paragraph.

For every comparison task, we create two balanced datasets, each with about 50% signal jets. The smaller one, named the sample dataset, consists a total of 10 000 jets and is mainly used for picking the best hyperparameters, though it also constitutes a complete analysis in its own right. The full dataset, on the other hand, has 140 000 jets in total and is used to assess the model performance and draw the final conclusions.

For the two classifiers kNN and SVM, the sample dataset is further divided into a training sample of 5000 jets, a validation sample of 2500 jets used to decide the best hyperparameters, and a test sample of 2500 jets. The full dataset is split into a training set of 100 000 jets and a test set of 40 000 jets for these two models. For LDA, thanks to its high efficiency, we train and test on both the sample dataset (training sample size equal to 8000, test sample size equal to 2000; validation sample is not needed since there is no hyperparameter for LDA) and the full dataset (training set size equal to 100 000, test set size equal to 40 000), which amounts to two separate, identical analyses. The k-medoids algorithm has only been applied to the sample dataset due to its computational intensity, and in this case, all 10 000 jets are fed into the model at once for clustering.

Figure 4 displays the receiver operating characteristic (ROC) curves of the three classifiers kNN, SVM, and LDA for each of the seven comparison tasks. Also included is the area under the ROC curve (AUC), which encapsulates the model performance in a single number between 0 and 1. An AUC close to 1 is most desirable, whereas a value around 0.5 suggests a random classifier, the worst-case scenario. All results are obtained on the full test datasets consisting of 40 000 jets, using the models trained on 100 000 jets with hyperparameters, if present, picked by the sample datasets.

To get a better sense of the model performance, we compare the AUCs of our LOT-coupled ML models for the W vs QCD classification task with other common classifiers built in Ref. [1] where the training set, though different, also contains 100 000 balanced W and QCD jets, and the test set contains 20 000 such jets. The model most akin to our $k_{=20}$NN-LOT is $k_{=32}$NN-EMD built upon the EMD proposed in Ref. [1], an interpolation between the OT-W1 distance and total variation norm.[5] The N-subjettiness ratio $\tau_2^{\beta=1}/\tau_1^{\beta=1}$, introduced in Refs. [41,42], is a widely used observable specifically designed to spot two-prong jet substructure. For the other three classifiers, namely the Energy Flow Network (EFN) and Particle Flow Network (PFN) neural networks [33], and a linear classier trained on Energy Flow Polynomials (EFPs) [32], please refer to the original papers for more details.

Datasets	Model	AUC
Our datasets	$k_{=20}$NN-LOT	0.845
	SVM-LOT	0.869
	LDA-LOT	0.704
Datasets in Ref. [1]	$k_{=32}$NN-EMD	0.887
	$\tau_2^{\beta=1}/\tau_1^{\beta=1}$	0.776
	PFN	0.919
	EFPs	0.917
	EFN	0.904

[5]Although our samples are not identical to those in Ref. [1], we apply the same prescription for simulating and preparing the samples, and our W/QCD jet samples yield results for $k_{=32}$NN-EMD compatible with Ref. [1].

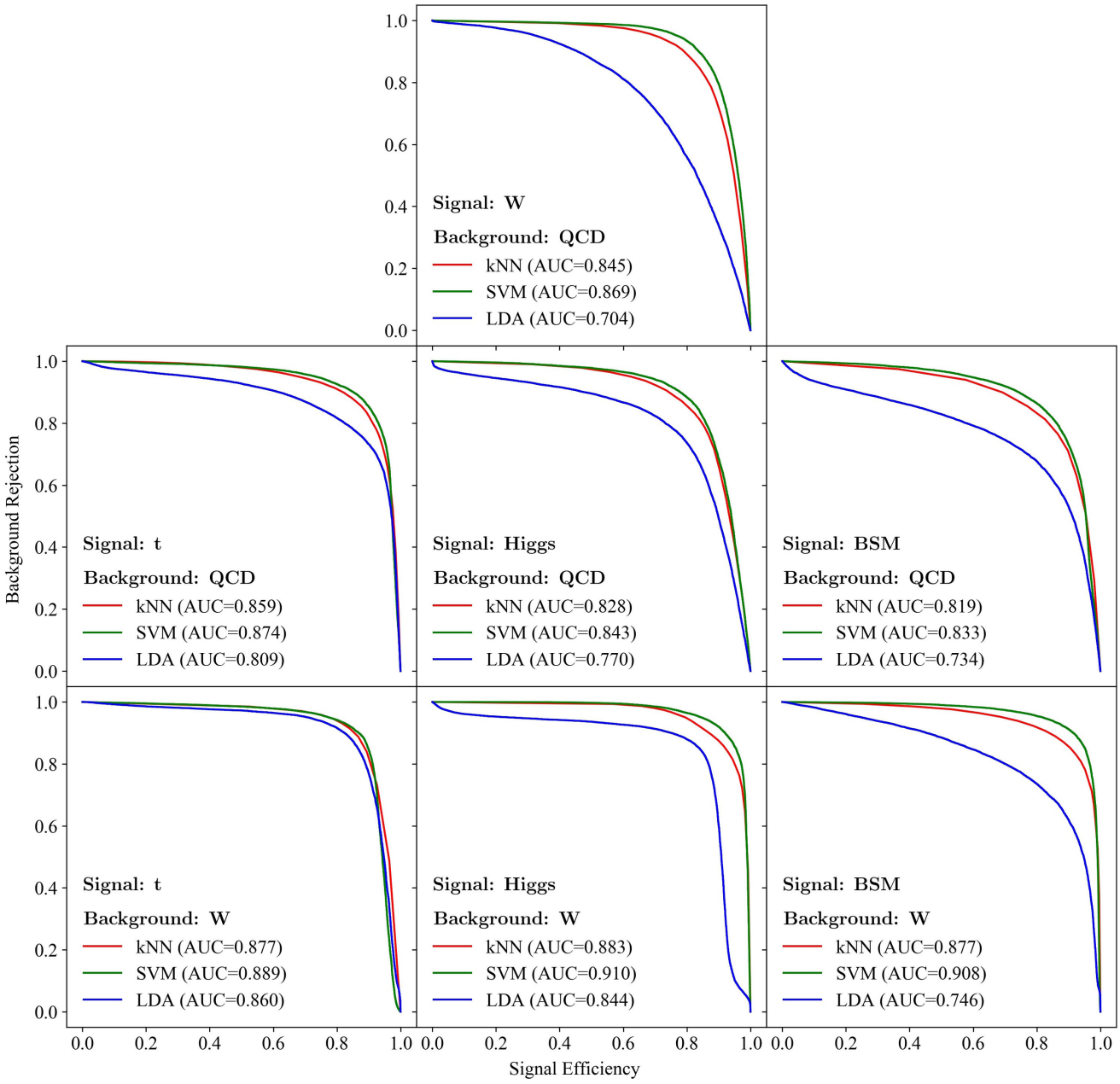

FIG. 4. ROC curves for the seven jet tagging tasks evaluated on the full test datasets of 40k jets. The x coordinate shows the signal efficiency rate and the y coordinate gives the background rejection rate.

Not surprisingly, the neural networks obtain the best performance. But the four optimal transport inspired models (three with LOT and one with EMD) are on a par with these state-of-the-art complex classifiers, and they significantly outperform the N-subjettiness observable (with the single exception of the exceptionally simplistic LDA). More pertinent to our current investigation is the observation that models coupled with LOT-W2 approximation perform as well as those using the exact EMD metric. The AUCs of kNN-LOT and SVM-LOT are close to the AUC of kNN-EMD, suggesting that it does not make much difference for jet tagging whether we use the exact OT metric or its linearized version. Yet on the practical level, the LOT approximation has a significant advantage over the exact OT metric. The computation of the LOT coordinates for 140 000 jets only takes about 10 min on a desktop computer, whereas it is infeasible to compute the full exact OT matrix of pairwise distances on the same computer and still requires significant time on a cluster.

Table I summarizes the results obtained for all seven comparison tasks, with complete, independent analyses done both on the sample datasets and the full datasets. In addition to AUC, we also report the true positive rate (TPR) and false positive rate (FPR), where the TPR is the same as

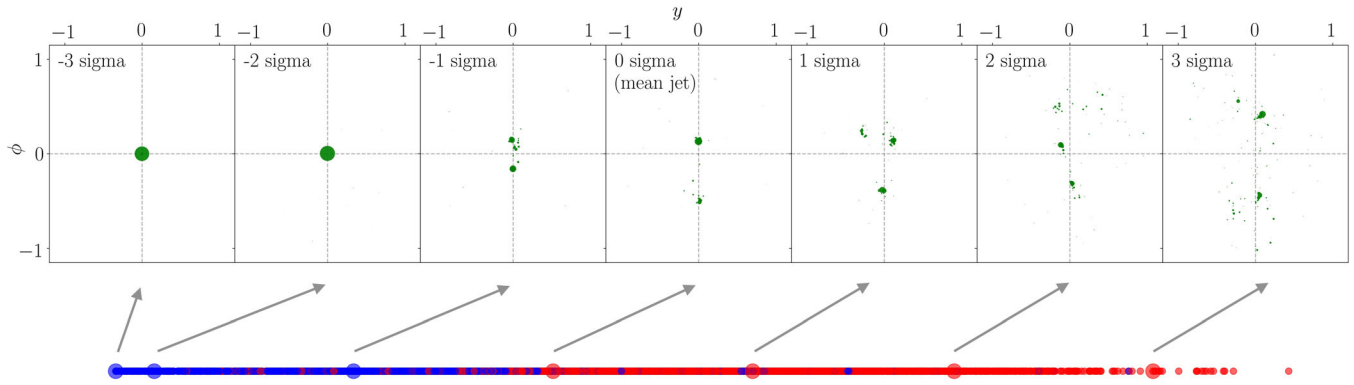

FIG. 5. Bottom: projection of the LOT coordinates of 10 000 jets in the sample dataset onto the LDA direction chosen by the model. Blue dots represent W boson jets and red dots refer to top jets. The seven larger dots represent jets whose LDA coordinates are $-3, -2, -1, 0, 1, 2, 3$ sigma away from the mean jet (starting from the left). Top: the energy flow in the rapidity-azimuthal plane of the seven jets chosen in the bottom plot respectively. The intersection of the dashed lines shows the location of the origin in the y-ϕ plane.

the signal efficiency and the FPR equals 1 minus the background rejection. A TPR near 1 and a FPR close to 0 are preferable. For SVM and kNN, we also include the hyperparameters chosen by the sample datasets. The results for k-medoids are harder to interpret, so we defer a full discussion to a later paragraph.

Also included in the table is the approximate run time for each task, performed on an iMac with 3.6 GHz 8-Core Intel Core i9 and 16 GB memory. The longest analysis takes no more than 10 h, which, when combined with the extra few minutes for calculating the LOT coordinates, is quite manageable. LDA in particular only takes seconds to process the full datasets, and in this light, its classification results are surprisingly good. In addition, models performed on the sample datasets require as few as 2 h for a full scan of hundreds of possible combinations of hyperparameters. Competitive classification performance coupled with efficient computational time suggests that the Linearized Optimal Transport metric may play a role in event classification alongside the exact OT metric, complex neural networks, and traditional handpicked observables.

Given that the sample datasets constitute complete analyses on their own rights, we can compare their results with those obtained using the full datasets. In general, model performance naturally gets better with more training data, but we observe that the increase in performance going from 10 000 jets to 140 000 jets is perhaps not significant enough to justify the extra computational resources needed. Since the numbers quoted for AUC, TPR, and FPR are only intended as general performance evaluations rather than precise measures, the fluctuations in these numbers can be safely ignored, and we therefore conclude that a dataset of 10 000 jets (with as few as 5000 for training) is already enough to assess the overall quality of the model and the underlying metric.

Some general features can be immediately read off from the table. Whichever jets we compare, SVM always gives the best classification performance with AUCs around 0.9, approaching the performance of neural networks. This suggests that jets represented in their LOT coordinates are indeed very well separated by a hyperplane in some high-dimensional feature space, which in turn demonstrates the fitness of the approximate metric itself. Except for t vs W jets classification, the hyperparameters chosen for SVM via the validation process are all the same, with $C = 1$ and $\gamma = 100$ where 1 happens to be the default value for C in SCIKIT-LEARN. It means that the model uses only a reasonable amount of regularization and thus a relatively smooth decision surface is drawn. On the other hand, a γ of 100 is considered large, indicating that only nearby samples can have an influence on the classification of a new point.

This latter observation is consistent with what is suggested by the hyperparameter k picked by kNN. All seven comparison tasks prefer small k values less than 50, which means that to determine the type of an unknown jet we need to look no further than its closest 50 neighbors. If LOT does not place same-type jets near each other as desired, then models with hyperparameters preferring locality will not be able to achieve such satisfying classification performances. Therefore, the hyperparameters picked by SVM and kNN provide an indirect evidence for the suitability of the optimal transport metric—it indeed groups jets of the same type near each other and separates those of different types. We will later turn this speculation into more convincing and intuitive visualization.

Among the seven jet tagging tasks, kNN and SVM both have the best performance in distinguishing Higgs boson jets from W boson jets and are least capable of separating BSM jets from QCD jets. This is mainly caused by a relatively high false positive rate, meaning that the models have a tendency to wrongly classify QCD jets as BSM jets. The same reason applies to LDA when it performs poorly on W vs QCD classification relative to other tasks. For each type of signal jets (t, H, or BSM), all three classification models perform better when the background is a W jet rather than a QCD jet.

We now focus on the k-medoids clustering algorithm, which is only analyzed on the sample datasets due to computational limitations. Given that unsupervised learning is inherently more difficult than supervised learning, it is not surprising to see the performance of k-medoids algorithm as inferior to that of kNN or SVM. But even then, except for the W vs QCD and BSM vs W tasks, the AUCs of k-medoids are all above 0.75, on a par with the supervised learning models analyzed on the sample datasets. The clustering algorithm even shows superior performance compared to LDA for most tagging tasks. This remarkable achievement again points to the merit of the underlying approximate LOT distance and is encouraging for the further exploration of optimal transport applications to unsupervised learning algorithms.

It should be noted that AUC is not the only gauge of model performance. Especially in the case of k-medoids clustering, we also need to take a look at other indicators to map a more complete picture. Beside examining the TPR and FPR, we also like to know more about the properties of the two clusters output by the algorithm. If the model is perfect, then each cluster should contain only signal jets or only background jets. The purity of the two clusters is given in the second row of k-medoids clustering in the table, where we record the signal percentage (defined as the number of signals in the cluster divided by the total number of jets in that cluster) in the signal cluster and the background cluster, respectively. By definition, the signal cluster is the group with a majority of signal jets, which, if pure, should have a signal percentage of 100%. Similarly, a pure background cluster should have 0% signal percentage. Notice that the sum of the signal percentage of the two clusters does not necessarily equal 1 (but in the ideal case, it does). The worst-case scenario is to have the signal percentage of both clusters close to 50%. A quick look at the second row at least qualitatively confirms that the AUC of the task is indeed higher whenever we have two purer clusters, with the best AUC obtained for t vs QCD clustering which has a signal percentage of 74.70% for the signal cluster and only 9.27% for the background cluster.

The size of the clusters also reveals how well the model performs. Ideally, the result would be two clusters with equal size, that is, each with 5000 jets, since the data themselves are balanced. Here, the best result we have is for the H vs QCD task, where the Higgs cluster has 5682 jets and the QCD cluster has a total of 4318 jets. But in general, the two clusters are not well balanced. In the worst case, the W cluster has 81.77% more jets than the BSM cluster, and it does correspond to the lowest AUC score.

In theory, the two medoids should be the most representative jet for the clusters to which they respectively belong. Since the medoids are actual data points, we can uncover their true labels and check whether they agree with the type of the cluster they are assigned to. Only the two tasks, t vs QCD and H vs W, give conflicting answers. For the t vs QCD clustering, the two chosen medoids are both background QCD jets. Thus, the signal top cluster acquires a QCD jet as its representative. The situation is reversed for the H vs W task where now the background W cluster elects a signal Higgs jet as its exemplar. Nevertheless, both tasks enjoy high AUC scores, which suggests that the true labels of the medoids might not have a direct influence on model performance.

The general message here is that AUC, though powerful and straightforward, is not enough to assess the performance of an algorithm; other indicators are required to gain a fuller appreciation of the strength and weakness of the model, both for clustering and for classification.

Lastly, we use LDA to visualize jets and aid understanding of the LOT approximation and its associated Euclidean embedding. Our approach follows work by Wang *et al.* [9], which introduced the LOT framework and applied it to visualization tasks, such as discriminating nuclear chromatin patterns in cancer cells. Given the 225×2 linearized coordinate for each jet, we first stack the list of the second coordinate ϕ at the end of the list of the first coordinate y and reshape the coordinate to be 450×1, which is then fed into a LDA model for the projection of the 450 coordinates onto one single most discriminative direction (denoted as the LDA direction). This allows us to represent every jet as one single point on the LDA direction for easy visualization. Figure 5 shows such projection for the 10 000 jets in the t vs W sample dataset, which enjoys the highest AUC among the seven tasks with the LDA classifier. A clear separation between W and top jets can be seen, with the majority of W boson jets grouped toward the left end of the LDA direction and most top jets towards the right end, explaining the good performance of the LDA classifier for this task.

It is enlightening to see how jets vary along the chosen LDA direction. To this end, we first select the jet whose one-dimensional projected LDA coordinate has a value closest to the mean of all LDA coordinates in the dataset and denote it as the mean jet. We then compute the standard deviation of the dataset. Now, jets whose LDA coordinates are up to 3 sigmas away from the mean jet are displayed in Fig. 5. We observe a clear tendency of particles spreading more on the y-ϕ plane as we move from the left end of the LDA direction to the right end, i.e., from negative sigmas to positive sigmas, corresponding well to our intuition that top jets are more smeared and tend to have a three-pronged structure.

As another illustration, we examine more closely how the OT-W2 metric rearranges the p_T of one jet to make it look like another, as shown in Fig. 6. Here, we first select the rightmost top jet t^1 and the leftmost W boson jet W^1 in the bottom plot of Fig. 5. We then compute the exact 2-Wasserstein optimal transportation matrix γ_{ij}, which instructs how much of p_T is moved from particle i in jet W^1 (denoted as W^1_i) to particle j in jet t^1 (denoted as t^1_j).

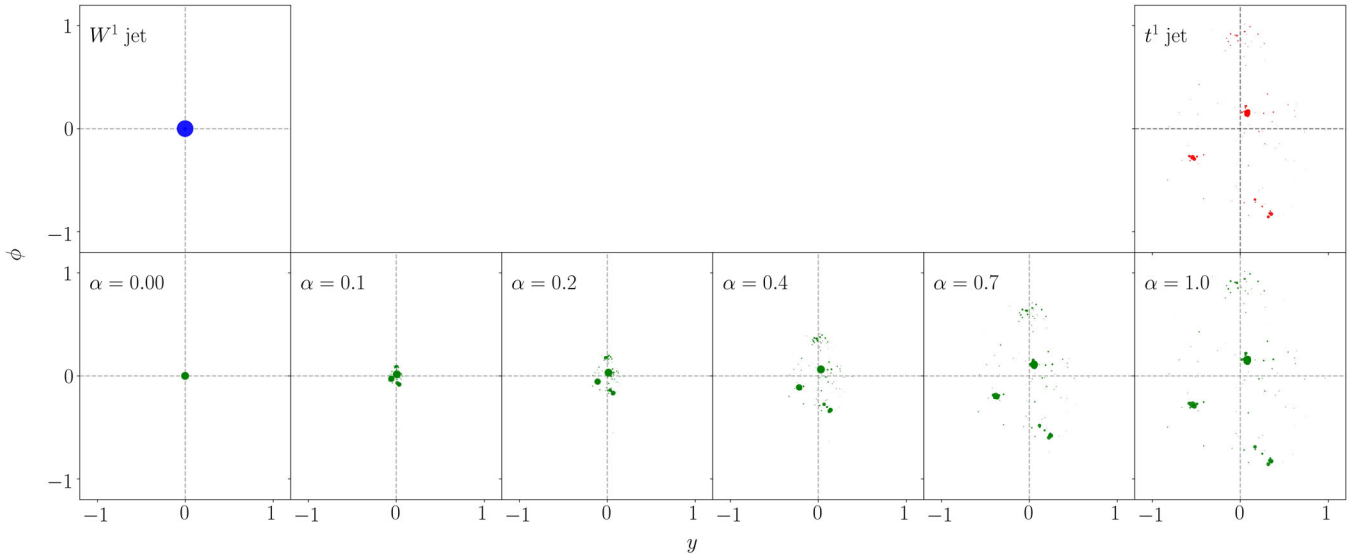

FIG. 6. The OT-W2 movement of p_T to rearrange the leftmost W boson jet W^1 (blue) into the rightmost top jet t^1 (red) in the sample dataset. The intermediate green plots show artificial jets created via the interpolation parameter α. When $\alpha = 0$ and 1, the jets are respectively identical to W^1 and t^1 up to visualization. Again, the intersection of the dashed lines shows the location of the origin.

To interpolate between the two extreme jets, we create a new jet that depends on an interpolation parameter $\alpha \in [0, 1]$, where $\alpha = 0$ outputs a jet identical to W^1 and $\alpha = 1$ recovers the t^1 jet. This new artificial jet$^\alpha$ contains $i \times j$ particles, each with

$$\begin{aligned} p_T^\alpha &= \gamma_{ij}, \\ y^\alpha &= (1-\alpha) \times y(W_i^1) + \alpha \times y(t_j^1), \\ \phi^\alpha &= (1-\alpha) \times \phi(W_i^1) + \alpha \times \phi(t_j^1), \end{aligned} \tag{7}$$

where $y(W_i^1)$ is the y coordinate of the ith particle in jet W^1, and likewise for the others. From the perspective of optimal transport theory, this artificial jet is precisely the 2-Wasserstein geodesic between the jets. Several values of α are picked in Fig. 6 so as to show a few representatives of the interpolated jets and help us understand intuitively the p_T movement by the OT-W2 metric. This interpolation technique may prove relevant to the fast simulation of collider events, insofar as it allows interpolation between real events.

The above visualizations provide useful insight into the performance of the LOT approximation and the machine learning model coupled to it, offering a useful intermediary between analytic kinematic variables and deep neural networks.

V. CONCLUSION

The theory of optimal transport offers a new perspective on the traditional problems of collider physics, beginning with the introduction of the OT-based Energy Mover's Distance in Ref. [1]. But the practical value of exact OT metrics as competitors to specialized variables and deep neural networks is limited by the need to determine $\mathcal{O}(N_{\text{evt}}^2)$ computationally expensive OT distances between N_{evt} events. In this paper, we have introduced an efficient approximation scheme for computing optimal transport distances in collider events using a linear optimal transport approximation to the 2-Wasserstein distance. This entails computing the exact OT distance between each event and a reference jet containing n particles; the corresponding transport plan provides a map from the event to a vector in n-dimensional Euclidean space. The approximate LOT distance between two events is then obtained by computing a simple weighted ℓ^2 distance between the corresponding n-vectors, so that only $\mathcal{O}(N_{\text{evt}})$ OT distances and $\mathcal{O}(N_{\text{evt}}^2)\ell^2$ distances are required. This makes the calculation of approximate OT distances between collider events in a typical sample accessible to a desktop computer. Furthermore, we have proven that this LOT approximation converges to a true metric on the space of collider events in the continuum limit.

The Euclidean embedding furnished by our approximation scheme makes it a natural input to simple machine learning algorithms that require more than the pairwise distance between events, such as LDA. We have demonstrated the value of the LOT framework for jet tagging in a number of classification tasks, illustrating both the relative computational efficiency (compared to exact OT approaches) and interpretability (compared to deep neural networks) of our approach. The two classifiers kNN and SVM coupled with the LOT approximation achieve high performance on a level comparable to both the exact OT approach and complex neural networks, while significantly outperforming the traditional N-subjettiness variable. The choice of the hyperparameters of the two models further

confirms the effectiveness of the approximate LOT distance in capturing the difference among various jet types. As a quick first look into the datasets, LDA performs surprisingly well and provides an intuitively clear visualization method. The good performance of the k-medoids clustering algorithm is encouraging for further explorations of the application of the LOT framework to tasks beyond supervised learning, including clustering and anomaly/novelty detection. Finally, the similarity in the performance of the sample datasets and the full datasets suggests that only as few as 10 000 jets are required to have an estimate on the quality of the model and the underlying metric, further reducing the computational cost.

There are a wide variety of future directions. The computational speedup offered by the LOT approximation should make it possible to apply optimal transport methods more broadly in analyzing both simulated and actual collider data. Likewise, this speedup motivates extending LOT methods to other optimal transport metrics (such as unbalanced OT) which may be relevant to collider physics but whose application is currently limited by computational cost. To the extent that it involves the transport plan from a reference jet to an event, the approximate LOT distance shares aspects with the OT-based event isotropy variable [4], and it would be interesting to investigate their relationship further. The convergence of the LOT approximation to a true metric in the continuum limit suggests it may play a role as a discrete approximation scheme in the broader geometric approach to collider observables proposed in Ref. [2].

More broadly, there remains much to explore at the interface between collider physics and the theory of optimal transport.

ACKNOWLEDGMENTS

J. C. would like to thank Timothy Trott for providing example FEYNRULES files. K. C. would like to thank Bernhard Schmitzer for helpful conversations about LOT. We would like to thank Eric Metodiev, Dejan Slepčev, and Jesse Thaler for comments on the manuscript. The work of T. C. and N. C. was supported in part by the Department of Energy under Grant No. DE-SC0011702. The work of J. C. was supported by The Create Fund, thanks to the generosity of CCS donors. The work of K. C. was supported by National Science Foundation Grant No. DMS-1811012 and a Hellman Faculty Fellowship.

APPENDIX: FROM LOT APPROXIMATION TO LOT DISTANCE

In this Appendix, we prove the convergence of the LOT approximation, defined in Eq. (4), to a true metric in the continuum limit. For the sake of brevity, we will only briefly discuss the optimal transport theory underlying this result, primarily with the goal of establishing notation. We refer the reader to the textbooks by Ambrogio *et al.* [43], Peyré and Cuturi [24], Santambrogio [44], and Villani [45] for further background.

Let $\mathcal{P}(\mathbb{R}^d)$ denote the set of probability measures on $\mathbb{R}^d$. Given $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, a measurable function $\mathbf{t}: \mathbb{R}^d \to \mathbb{R}^d$ *transports* μ *onto* ν if $\nu(B) = \mu(\mathbf{t}^{-1}(B))$ for all measureable sets $B \subseteq \mathbb{R}^d$. We call ν the *push-forward of* μ *under* $\mathbf{t}$ and write $\nu = \mathbf{t}\#\mu$. For historical reasons, it is conventional in the field of optimal transport to think of the amount of measure μ gives to a measurable set B as the *mass of* B *with respect to* μ and to interpret a measurable function $\mathbf{t}$ as a *transport map* that *rearranges the mass in* μ *to look like* ν. Conveniently for physicists, the "mass" and "energy" notation is equivalent in natural units, and we will use the former here. Given a probability measure on a product space, for example $\boldsymbol{\gamma} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$, its *marginals* are given by the push-forward of the measure through the projections on each component of the product. For example, if $\pi^2: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is the projection onto the second component of $\mathbb{R}^d \times \mathbb{R}^d$, then $\pi^2\#\boldsymbol{\gamma}$ is the *second marginal* of γ. Finally, we say that $\mathcal{E} \in \mathcal{P}(\mathbb{R}^d)$ has *finite second moment* if $M_2(\mathcal{E}) \coloneqq \int_{\mathbb{R}^d} |x|^2 d\mathcal{E}(x) < +\infty$, in which case we write $\mathcal{E} \in \mathcal{P}_2(\mathbb{R}^d)$.

For any $\mathcal{E}, \tilde{\mathcal{E}} \in P_2(\mathbb{R}^d)$, the 2-Wasserstein distance from $\mathcal{E}$ to $\tilde{\mathcal{E}}$ is given by

$$W_2(\mathcal{E}, \tilde{\mathcal{E}}) = \min_{\boldsymbol{\gamma} \in \Gamma(\mathcal{E}, \tilde{\mathcal{E}})} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - y|^2 d\boldsymbol{\gamma}(x, y) \right)^{1/2},$$

$$\Gamma(\mathcal{E}, \tilde{\mathcal{E}}) = \{\boldsymbol{\gamma} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : \pi^1\#\boldsymbol{\gamma} = \mathcal{E}, \pi^2\#\boldsymbol{\gamma} = \tilde{\mathcal{E}}\}.$$

Note that, in the special case $\mathcal{E} = \sum_i \delta_{x_i} E_i$, $\tilde{\mathcal{E}} = \sum_j \delta_{\tilde{x}_j} \tilde{E}_j$, the above definition of the 2-Wasserstein distance coincides with that given in Sec. II. We refer to the set of transport plans $\gamma \in \Gamma(\mathcal{E}, \tilde{\mathcal{E}})$ that achieve the minimum as the set of *optimal transport plans*, which we denote by $\Gamma_0(\mu, \nu)$. Furthermore, we say that a plan $\boldsymbol{\gamma} \in \Gamma(\mathcal{E}, \tilde{\mathcal{E}})$ is *induced by a transport map* if there exists a measurable function $\mathbf{t}: \mathbb{R}^d \to \mathbb{R}^d$ so that $\boldsymbol{\gamma} = (\mathbf{id} \times \mathbf{t})\#\mathcal{E}$, where $\mathbf{id}(x) = x$ is the identity mapping.

Just as we may extend the 2-Wasserstein distance from the discrete case to the case of probability measures, we may likewise extend the definition of the LOT functional, as well as define the related concept of transport metrics. We devote particular attention to the case that the reference measure $\mathcal{R}$ does not give mass to sets of $(d-1)$-dimensional Hausdorff measure; in other words, the measure does not concentrate on small sets. In this case, for any $\mathcal{E} \in \mathcal{P}_2(\mathbb{R}^d)$, there exists a unique optimal transport plan $\boldsymbol{\rho} \in \Gamma_0(\mathcal{R}, \mathcal{E})$, and $\boldsymbol{\rho}$ is induced by a transport map [46]. This transport map is unique (up to sets of $\mathcal{E}$ measure zero), and we refer to it as the *optimal transport map from* $\mathcal{R}$ *to* $\mathcal{E}$, denoted $\mathbf{t}_{\mathcal{R}}^{\mathcal{E}}$ [47]. The function $x \mapsto \mathbf{t}_{\mathcal{R}}^{\mathcal{E}}(x)$ represents where mass starting at location x in the reference measure $\mathcal{R}$ is

sent in the target measure $\mathcal{E}$, in order to rearrange the mass from $\mathcal{R}$ into $\mathcal{E}$, using the least amount of effort. Note that a necessary condition for such an optimal transport map to exist is that an optimal rearrangement of $\mathcal{R}$ to $\mathcal{E}$ does not *split mass*; that is, all mass starting at a specific location in $\mathcal{R}$ must be sent to the same location in $\mathcal{E}$.

Given a reference measure $\mathcal{R} \in \mathcal{P}_2(\mathbb{R}^d)$, which does not give mass to sets of $(d-1)$-dimensional Hausdorff measure, and measures $\mathcal{E}, \tilde{\mathcal{E}} \in \mathcal{P}_2(\mathbb{R}^d)$, the *transport metric with base* $\mathcal{R}$ is given by

$$W_{2,\mathcal{R}}(\mathcal{E}, \tilde{\mathcal{E}}) = \left(\int |\mathbf{t}_{\mathcal{R}}^{\mathcal{E}} - \mathbf{t}_{\mathcal{R}}^{\tilde{\mathcal{E}}}|^2 d\mathcal{R} \right)^{1/2}. \tag{A1}$$

The transport metric with base $\mathcal{R}$ is a well-defined metric on $\mathcal{P}_2(\mathbb{R}^d)$, which can be interpreted as computing the distance between $\mathcal{E}$ and $\tilde{\mathcal{E}}$ by projecting onto the tangent plane at $\mathcal{R}$ (see Ref. [48], Proposition 1.15, Ref. [9], Eq. (6), and Ref. [43], Eqs. (7.3.2) and (9.2.5) and Theorem 8.5.1).

In this section, we prove that the Linearized Optimal Transport approximation converges as the discretization of the reference measure is refined. In order to do this, we now define the LOT functional for general measures and show its relationship with the transport metric with base $\mathcal{R}$. Given measures $\mathcal{R}, \mathcal{E}, \tilde{\mathcal{E}} \in \mathcal{P}_2(\mathbb{R}^d)$, for any $\boldsymbol{\rho} \in \Gamma_0(\mathcal{R}, \mathcal{E})$, $\tilde{\boldsymbol{\rho}} \in \Gamma_0(\mathcal{R}, \tilde{\mathcal{E}})$, there exists $\boldsymbol{\omega} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$ so that

$$\boldsymbol{\pi}^{1,2}\#\boldsymbol{\omega} = \boldsymbol{\rho} \quad \text{and} \quad \boldsymbol{\pi}^{1,3}\#\boldsymbol{\omega} = \tilde{\boldsymbol{\rho}}, \tag{A2}$$

where $\boldsymbol{\pi}^{i,j}$ is the projection on the ith and jth components of $\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$; when $\mathcal{R}$ does not give mass to small sets; then, $\boldsymbol{\omega}$ is unique (see Ref. [43], Lemma 5.3.2). By disintegration of measures, there exists a family $\{\boldsymbol{\omega}_{x_1} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)\}_{x_1 \in \mathbb{R}^d}$ so that for any measurable function $f : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \to [0, +\infty)$,

$$\int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} f(x_1, x_2, x_3) d\boldsymbol{\omega}_{x_1}(x_2, x_3) \right) d\mathcal{R}(x_1) = \int_{\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d} f(x_1, x_2, x_3) d\boldsymbol{\omega}(x_1, x_2, x_3). \tag{A3}$$

In this way, for $\mathcal{R}, \mathcal{E}, \tilde{\mathcal{E}} \in \mathcal{P}_2(\mathbb{R}^d)$ and $\boldsymbol{\rho} \in \Gamma_0(\mathcal{R}, \mathcal{E})$, $\tilde{\boldsymbol{\rho}} \in \Gamma_0(\mathcal{R}, \tilde{\mathcal{E}})$, the LOT functional is defined by

$$LOT_{\boldsymbol{\rho}, \tilde{\boldsymbol{\rho}}}(\mathcal{E}, \tilde{\mathcal{E}}) = \left(\int \left| \int (x_2 - x_3) d\boldsymbol{\omega}_{x_1}(x_2, x_3) \right|^2 d\mathcal{R}(x_1) \right)^{1/2}. \tag{A4}$$

In the special case that $\mathcal{R} = \sum_i \delta_{y_i} R_i$, $\mathcal{E} = \sum_j \delta_{x_j} E_j$, and $\tilde{\mathcal{E}} = \sum_k \delta_{\tilde{x}_k} \tilde{E}_k$, this reduces to the LOT functional defined in Sec. II. Furthermore, in the special case that $\mathcal{R}$ does not give mass to sets of $(d-1)$-dimensional Hausdorff measure, the optimal transport plans $\boldsymbol{\rho} = (\mathbf{id} \times \mathbf{t}_{\mathcal{R}}^{\mathcal{E}})\#\mathcal{R}$ and $\tilde{\boldsymbol{\rho}} = (\mathbf{id} \times \mathbf{t}_{\mathcal{R}}^{\tilde{\mathcal{E}}})\#\mathcal{R}$ are unique, as is the measure $\boldsymbol{\omega} = (\mathbf{id} \times \mathbf{t}_{\mathcal{R}}^{\mathcal{E}} \times \mathbf{t}_{\mathcal{R}}^{\tilde{\mathcal{E}}})\#\mathcal{R}$ and its disintegration $\boldsymbol{\omega}_{x_1} = \delta_{(\mathbf{t}_{\mathcal{R}}^{\mathcal{E}}(x_2), \mathbf{t}_{\mathcal{R}}^{\tilde{\mathcal{E}}}(x_3))}$. Consequently, when $\mathcal{R}$ does not give mass to small sets, the LOT functional is independent of the choice of transport plans $\boldsymbol{\rho}, \tilde{\boldsymbol{\rho}}$, and $\mathrm{LOT}_{\boldsymbol{\rho}, \tilde{\boldsymbol{\rho}}}(\mathcal{E}, \tilde{\mathcal{E}}) = W_{2,\mathcal{R}}(\mathcal{E}, \tilde{\mathcal{E}})$; that is, the LOT approximation becomes a well-defined metric on the space of probability measures with finite second moment. Similarly, when $\mathcal{R}$ does not give mass to small sets, the LOT Euclidean embedding can be thought of, from a geometric perspective, as the inverse of the exponential map

$$\mathcal{E} \mapsto \int x_2 d\boldsymbol{\omega}_{x_1}(x_2, x_3) = \mathbf{t}_{\mathcal{R}}^{\mathcal{E}}, \tag{A5}$$

which is an isometric embedding from $W_{2,\mathcal{R}}$ to $L^2(\mathcal{R})$.

We now prove that, for any sequence $\mathcal{R}^N \xrightarrow{W_2} \mathcal{R}$, where $\mathcal{R}$ does not give mass to small sets, the LOT approximation corresponding to $\mathcal{R}^N$ converges to the transport metric with base $\mathcal{R}$. Furthermore, we allow the events $\mathcal{E}^N$ and $\tilde{\mathcal{E}}^N$ to likewise vary along convergent sequences.

Proposition 1: Consider three sequences of probability measures $\mathcal{R}^N, \mathcal{E}^N, \tilde{\mathcal{E}}^N \in \mathcal{P}_2(\mathbb{R}^d)$ that converge to $\mathcal{R}$, $\mathcal{E}$, and $\tilde{\mathcal{E}}$ in the 2-Wasserstein metric. If $\mathcal{R}$ does not give mass to small sets, then for any choices of optimal transport plans $\boldsymbol{\rho}^N \in \Gamma_0(\mathcal{R}^N, \mathcal{E}^N)$ and $\tilde{\boldsymbol{\rho}}^N \in \Gamma_0(\mathcal{R}^N, \tilde{\mathcal{E}}^N)$, we have

$$\lim_{N \to +\infty} LOT_{\boldsymbol{\rho}^N, \tilde{\boldsymbol{\rho}}^N}(\mathcal{E}^N, \tilde{\mathcal{E}}^N) = W_{2,\mathcal{R}}(\mathcal{E}, \tilde{\mathcal{E}}). \tag{A6}$$

Proof.—Throughout, we use the equivalence between convergence in the Wasserstein metric and narrow convergence combined with convergence of second moments (see Ref. [43], Remark 7.1.11). In particular, this fact ensures that $\mathcal{R}^N$, $\mathcal{E}^N$, and $\tilde{\mathcal{E}}^N$ converge narrowly, so $\boldsymbol{\omega}^N$ is narrowly relatively compact (see Ref. [43], Lemma 5.2.2). Any narrow limit point $\boldsymbol{\omega}$ of this sequence satisfies, in the sense of narrow convergence,

$$\boldsymbol{\pi}^{1,2}\#\boldsymbol{\omega} = \lim_{N \to +\infty} \boldsymbol{\pi}^{1,2}\#\boldsymbol{\omega}^N = \lim_{N \to +\infty} \boldsymbol{\rho}^N = \boldsymbol{\rho}, \tag{A7}$$

$$\boldsymbol{\pi}^{1,3}\#\boldsymbol{\omega} = \lim_{N \to +\infty} \boldsymbol{\pi}^{1,3}\#\boldsymbol{\omega}^N = \lim_{N \to +\infty} \tilde{\boldsymbol{\rho}}^N = \tilde{\boldsymbol{\rho}}, \tag{A8}$$

where $\boldsymbol{\rho} \in \Gamma_0(\mathcal{R}, \mathcal{E})$, $\tilde{\boldsymbol{\rho}} \in \Gamma_0(\mathcal{R}, \tilde{\mathcal{E}})$ (see Ref. [43], Proposition 7.1.3). Since $\mathcal{R}$ does not give mass to sets of $(d-1)$-dimensional Hausdorff measure, the limit point $\boldsymbol{\omega}$ is unique, and $\boldsymbol{\omega} = (\mathbf{id} \times \mathbf{t}_{\mathcal{R}}^{\mathcal{E}} \times \mathbf{t}_{\mathcal{R}}^{\tilde{\mathcal{E}}})\#\mathcal{R}$ (see Ref. [43], Lemma 5.3.2). Furthermore, since

$$\begin{aligned}
&\lim_{N\to+\infty} M_2(\boldsymbol{\omega}^N)\\
&\quad= \lim_{N\to+\infty} \int |x_1|^2+|x_2|^2+|x_3|^2 d\boldsymbol{\omega}^N(x_1,x_2,x_3)\\
&\quad= \lim_{N\to+\infty} M_2(\mathcal{R}^N)+M_2(\mathcal{E}^N)+M_2(\tilde{\mathcal{E}}^N)\\
&\quad= M_2(\mathcal{R})+M_2(\mathcal{E})+M_2(\tilde{\mathcal{E}})=M_2(\boldsymbol{\omega}),
\end{aligned} \tag{A9}$$

we obtain that $\boldsymbol{\omega}^N \to \boldsymbol{\omega}$ not only narrowly but also in the Wasserstein metric.

We now apply this convergence of $\boldsymbol{\omega}^N$ to $\boldsymbol{\omega}$ to conclude the convergence of the LOT approximation to the transport metric with base $\mathcal{R}$. First, we will show

$$\limsup_{N\to+\infty} LOT_{\boldsymbol{\rho}^N,\tilde{\boldsymbol{\rho}}^N}(\mathcal{E}^N,\tilde{\mathcal{E}}^N) \le W_{2,\mathcal{R}}(\mathcal{E},\tilde{\mathcal{E}}). \tag{A10}$$

By Jensen's inequality for the probability measures $\boldsymbol{\omega}^N_{x_1}$,

$$\begin{aligned}
&LOT_{\boldsymbol{\rho}^N,\tilde{\boldsymbol{\rho}}^N}(\mathcal{E}^N,\tilde{\mathcal{E}}^N)\\
&\quad\le \left(\iint |x_2-x_3|^2 d\boldsymbol{\omega}^N_{x_1}(x_2,x_3) d\mathcal{R}^N(x_1)\right)^{1/2}\\
&\quad= \left(\int |x_2-x_3|^2 d\boldsymbol{\omega}^N(x_1,x_2,x_3)\right)^{1/2}.
\end{aligned} \tag{A11}$$

Taking the limsup as $N\to+\infty$ and using the convergence of $\boldsymbol{\omega}^N$ to $\boldsymbol{\omega}=(\mathbf{id}\times\mathbf{t}^{\mathcal{E}}_{\mathcal{R}}\times\mathbf{t}^{\tilde{\mathcal{E}}}_{\mathcal{R}})\#\mathcal{R}$ in the Wasserstein metric gives inequality (A10) (see Ref. [43], Lemma 5.1.7 and Proposition 7.1.5).

It remains to show that

$$\liminf_{N\to+\infty} \mathrm{LOT}_{\boldsymbol{\rho}^N,\tilde{\boldsymbol{\rho}}^N}(\mathcal{E}^N,\tilde{\mathcal{E}}^N) \ge W_{2,\mathcal{R}}(\mathcal{E},\tilde{\mathcal{E}}). \tag{A12}$$

Since $\mathcal{R}$ does not give mass to sets of $(d-1)$-dimensional Hausdorff measure, $W_{2,\mathcal{R}}(\mathcal{E},\tilde{\mathcal{E}})=LOT_{\boldsymbol{\rho},\boldsymbol{\rho}'}(\mathcal{E},\tilde{\mathcal{E}})$, and, squaring both sides, it is equivalent to show

$$\liminf_{N\to+\infty} \int |v^N(x_1)|^2 d\mathcal{R}^N(x_1) \ge \int |v(x_1)|^2 d\mathcal{R}(x_1), \tag{A13}$$

where

$$\begin{aligned}
v^N(x_1) &= \int (x_2-x_3) d\boldsymbol{\omega}^N_{x_1}(x_2,x_3)\\
v(x_1) &= \int (x_2-x_3) d\boldsymbol{\omega}_{x_1}(x_2,x_3)
\end{aligned} \tag{A14}$$

Since $\mathcal{R}^N\to\mathcal{R}$ narrowly and $x\mapsto|x|^2$ is convex, this holds as long as $v^N\in L^2(\mathcal{R}^N)$ weakly converge to $v\in L^2(\mathcal{R})$ (see Ref. [43], Theorem 5.4.4 (ii)). Indeed, for any $f\in C^\infty_c(\mathbb{R}^d)$, the fact that $\boldsymbol{\omega}^N\to\boldsymbol{\omega}$ in the Wasserstein metric ensures

$$\begin{aligned}
&\lim_{N\to+\infty}\int f(x_1)v^N(x_1)d\mathcal{R}^N(x_1)\\
&\quad= \lim_{N\to+\infty}\iint f(x_1)(x_2-x_3)d\boldsymbol{\omega}^N_{x_1}(x_2,x_3)d\mathcal{R}^N(x_1)\\
&\quad= \lim_{N\to+\infty}\int f(x_1)(x_2-x_3)d\boldsymbol{\omega}^N(x_1,x_2,x_3)\\
&\quad= \int f(x_1)(x_2-x_3)d\boldsymbol{\omega}(x_1,x_2,x_3)\\
&\quad= \iint f(x_1)(x_2-x_3)d\boldsymbol{\omega}_{x_1}(x_2,x_3)d\mathcal{R}(x_1)\\
&\quad= \int f(x_1)v(x_1)d\mathcal{R}(x_1).
\end{aligned} \tag{A15}$$

■

Corollary 3: Let Ω be a two-dimensional rectangular domain, and consider a sequence of reference measures $\mathcal{R}^N$ given by a sum of N^2 Dirac masses with weights $1/N^2$, uniformly distributed on Ω. Then, as $N\to+\infty$, the LOT approximation with base $\mathcal{R}^N$ converges to the transport metric with base $\mathcal{R}$, where $\mathcal{R}$ is the probability measure uniformly distributed on Ω. That is, for any events $\mathcal{E},\tilde{\mathcal{E}}$, and for any $\boldsymbol{\rho}\in\Gamma_0(\mathcal{R}^N,\mathcal{E})$, $\tilde{\boldsymbol{\rho}}\in\Gamma_0(\mathcal{R}^N,\tilde{\mathcal{E}})$, we have

$$\lim_{N\to_\infty} \mathrm{LOT}_{\boldsymbol{\rho}^N,\tilde{\boldsymbol{\rho}}^N}(\mathcal{E},\tilde{\mathcal{E}}) = W_{2,\mathcal{R}}(\mathcal{E},\tilde{\mathcal{E}}). \tag{A16}$$

Proof.—Note that, by construction, $\mathcal{R}^N$ converges in the Wasserstein metric to the probability measure uniformly distributed on Ω, which does not give mass to small sets. Consequently, the result follows from Proposition 1. ■

[1] P. T. Komiske, E. M. Metodiev, and J. Thaler, Metric Space of Collider Events, Phys. Rev. Lett. **123**, 041801 (2019).

[2] P. T. Komiske, E. M. Metodiev, and J. Thaler, The hidden geometry of particle collisions, J. High Energy Phys. 07 (2020) 006.

[3] P. T. Komiske, R. Mastandrea, E. M. Metodiev, P. Naik, and J. Thaler, Exploring the space of jets with CMS open data, Phys. Rev. D **101**, 034009 (2020).

[4] C. Cesarotti and J. Thaler, A robust measure of event isotropy at colliders, J. High Energy Phys. 08 (2020) 084.

[5] T. Cheng, J.-F. Arguin, J. Leissner-Martin, J. Pilette, and T. Golling, Variational autoencoders for anomalous jet tagging, arXiv:2007.01850.
[6] M. C. Romao, N. Castro, J. Milhano, R. Pedro, and T. Vale, Use of a generalized energy Mover's distance in the search for rare phenomena at colliders, arXiv:2004.09360.
[7] A. Mullin, H. Pacey, M. Parker, M. White, and S. Williams, Does SUSY have friends? A new approach for LHC event analysis, arXiv:1912.10625.
[8] A. Y. Wei, P. Naik, A. W. Harrow, and J. Thaler, Quantum algorithms for jet clustering, Phys. Rev. D **101**, 094015 (2020).
[9] W. Wang, D. Slepčev, S. Basu, J. A. Ozolek, and G. K. Rohde, A linear optimal transportation framework for quantifying and visualizing variations in sets of images, Int. J. Comput. Vis. **101**, 254 (2013).
[10] P. T. Komiske, E. M. Metodiev, and J. Thaler, Cutting multiparticle correlators down to size, Phys. Rev. D **101**, 036019 (2020).
[11] L. G. Hanin, Kantorovich-Rubinstein norm and its application in the theory of Lipschitz spaces, Proc. Am. Math. Soc. **115**, 345 (1992).
[12] B. Piccoli and F. Rossi, Generalized Wasserstein distance and its application to transport equations with source, Arch. Ration. Mech. Anal. **211**, 335 (2014).
[13] B. Piccoli and F. Rossi, On properties of the generalized Wasserstein distance, Arch. Ration. Mech. Anal. **222**, 1339 (2016).
[14] M. Thorpe, S. Park, S. Kolouri, G. K. Rohde, and D. Slepčev, A transportation l^p distance for signal analysis, J. Math. Imaging Vision **59**, 187 (2017).
[15] O. Pele and M. Werman, A linear time histogram metric for improved sift matching, in *European Conference on Computer Vision* (Springer, New York, 2008), pp. 495–508.
[16] O. Pele and M. Werman, Fast and robust earth Mover's distances, in *2009 IEEE 12th International Conference on Computer Vision* (IEEE, Kyoto,2009), pp. 460–467.
[17] Y. Rubner, C. Tomasi, and L. J. Guibas, The earth Mover's distance as a metric for image retrieval, Int. J. Comput. Vis. **40**, 99 (2000).
[18] W. Wang, J. A. Ozolek, D. Slepčev, A. B. Lee, C. Chen, and G. K. Rohde, An optimal transportation approach for nuclear structure-based pathology, IEEE Trans. Med. Imaging **30**, 621 (2010).
[19] J. Delon, Midway image equalization, J. Math. Imaging Vision **21**, 119 (2004).
[20] J. Altschuler, J. Niles-Weed, and P. Rigollet, Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration, in *Advances in Neural Information Processing Systems* (2017), pp. 1964–1974.
[21] D. P. Bertsekas, A new algorithm for the assignment problem, Math. Program. **21**, 152 (1981).
[22] D. P. Bertsekas and J. Eckstein, Dual coordinate step methods for linear network flow problems, Math. Program. **42**, 203 (1988).
[23] M. Cuturi, Sinkhorn distances: Lightspeed computation of optimal transport, in *Advances in Neural Information Processing Systems* (2013), pp. 2292–2300.
[24] G. Peyré, M. Cuturi *et al.*, Computational optimal transport: With applications to data science, Found. Trends Mach. Learn. **11**, 355 (2019).
[25] Q. Mérigot, A. Delalande, and F. Chazal, Quantitative stability of optimal transport maps and linearization of the 2-wasserstein space, in *International Conference on Artificial Intelligence and Statistics* (2020), pp. 3186–3196.
[26] V. Seguy and M. Cuturi, Principal geodesic analysis for probability measures under the optimal transport metric, in *Advances in Neural Information Processing Systems* (2015), pp. 3312–3320.
[27] A. Aldroubi, S. Li, and G. K. Rohde, Partitioning signal classes using transport transforms for data analysis and machine learning, arXiv:2008.03452.
[28] C. Moosmüller and A. Cloninger, Linear optimal transport embedding: Provable fast Wasserstein distance computation and classification for nonlinear problems, arXiv:2008.09165.
[29] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations, J. High Energy Phys. 07 (2014) 079.
[30] T. Sjstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, An introduction to pythia 8.2, Comput. Phys. Commun. **191**, 159177 (2015).
[31] M. Cacciari, G. P. Salam, and G. Soyez, Fastjet user manual, Eur. Phys. J. C **72**, 1896 (2012).
[32] P. T. Komiske, E. M. Metodiev, and J. Thaler, Energy flow polynomials: A complete linear basis for jet substructure, J. High Energy Phys. 04 (2018) 013.
[33] P. T. Komiske, E. M. Metodiev, and J. Thaler, Energy flow networks: Deep sets for particle jets, J. High Energy Phys. 01 (2019) 121.
[34] R. Flamary and N. Courty, POT Python Optimal Transport library, https://pythonot.github.io/ (2017).
[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011).
[36] T. Cover and P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory **13**, 21 (1967).
[37] C. Cortes and V. N. Vapnik, Support-vector networks, Mach. Learn. **20**, 273 (1995).
[38] R. A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugenics **7**, 179 (1936).
[39] L. Kaufman and P. Rousseeuw, Clustering by means of medoids, in Statistical Data Analysis Based on the L_1 Norm and Related Methods, edited by Y. Dodge (Springer, New York, 1987), pp. 405–416.
[40] A. V. Novikov, Pyclustering: Data mining library, J. Open Source Softw. **4**, 1230 (2019).
[41] J. Thaler and K. Van Tilburg, Identifying boosted objects with n-subjettiness, J. High Energy Phys. 03 (2011) 015.

[42] J. Thaler and K. Van Tilburg, Maximizing boosted top identification by minimizing n-subjettiness, J. High Energy Phys. 02 (2012) 093.
[43] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient Flows: In Metric Spaces and in the Space of Probability Measures* (Springer Science & Business Media, New York, 2008).
[44] F. Santambrogio, in *Optimal Transport for Applied Mathematicians* (Birkäuser, New York, 2015), Vol. 55, p. 94.
[45] C. Villani, *Topics in Optimal Transportation*, Vol. 58 (American Mathematical Society, Providence, 2003).
[46] N. Gigli, On the inverse implication of Brenier-McCann theorems and the structure of $(P_2(M), W_2)$, Methods Appl. Anal. **18**, 127 (2011).
[47] R. J. McCann, Existence and uniqueness of monotone measure-preserving maps, Duke Math. J. **80**, 309 (1995).
[48] K. Craig, The exponential formula for the Wasserstein metric, ESAIM Control Optim. Calc. Var. **22**, 169 (2016).