

May 9, 2022

Session: QCD&EW I

1

Optimal Transport for Jet Physics

Speaker: Tianji Cai

**Conference: Phenomenology 2022 Symposium
(University of Pittsburgh)**

Based on **“Which metric on the space of collider events?”** [2111.03670]

“Linearized Optimal Transport for Collider Events” [2008.08604]

w/ Junyi Cheng, Nathaniel Craig (P.I.) & Katy Craig,

“The Linearized Hellinger-Kantorovich Distance” [2102.08807]

w/ J. Cheng, B. Schmitzer, M. Thorpe

Contents

- **Introduction:**

Why OT & What is OT, OT as the Metric, EMD & Its Generalization

- **Theory of Optimal transport:**

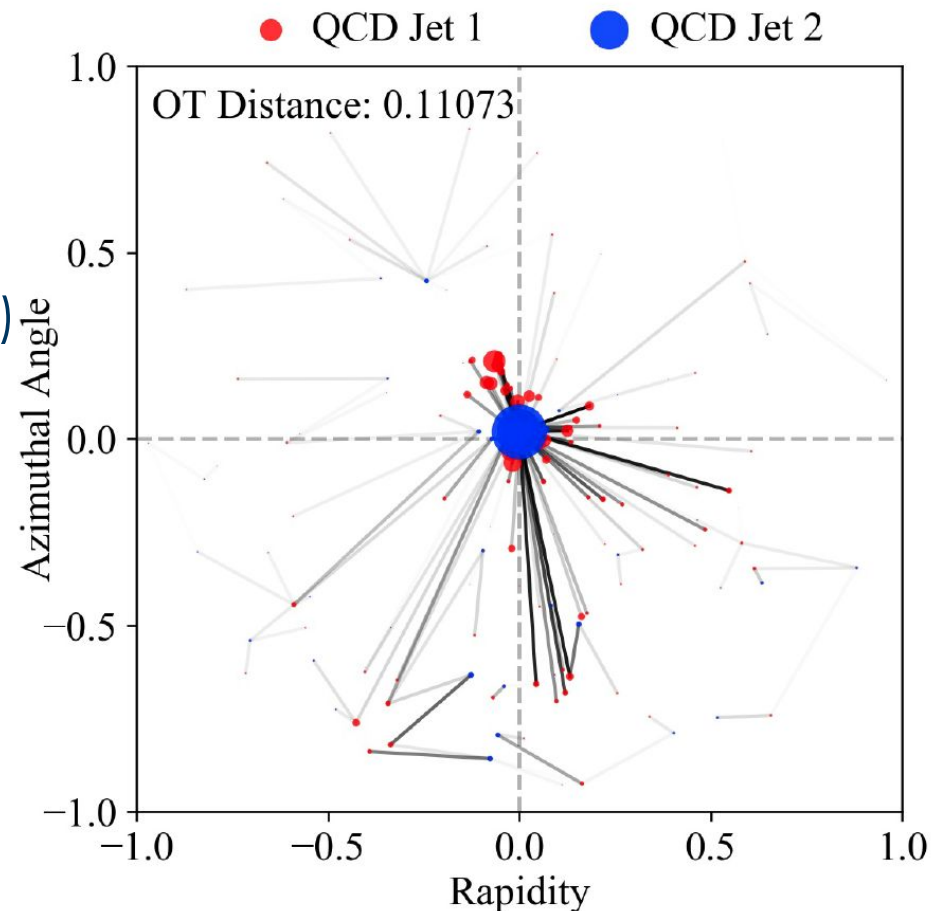
Balanced OT, Unbalanced OT,

Practical Limitation,

Linearized Optimal Transport (LOT)

- **LOT for Jet Tagging**

- **Summary**



1. Introduction: Why Optimal Transport & What is OT?

Goal: Want a way to quantify the **distance** between collider events/jets.



Optimal Transport is a well-developed mathematical theory defining a family of metrics between two distributions.

1781: Gaspard Monge,
Mémoire sur la théorie des déblais et des remblais
(*On cuttings and embankments*)



1942: Leonid Kantorovich,
On the translocation of masses



1999: Felix Otto,
The geometry of dissipative evolution equations: the porous medium equation



2000: Felix Otto, Cedric Villani,
Generalization of an inequality by Talagrand, as a consequence of the logarithmic Sobolev inequality



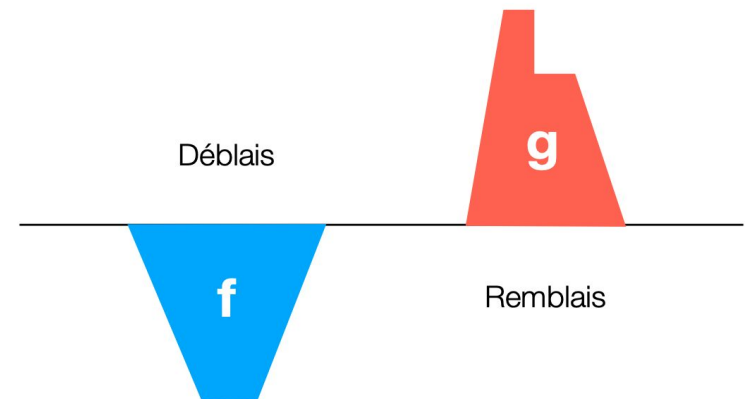
2010: Cedric Villani wins Fields medal

2018: Alessio Figalli wins Fields medal



Fundamental problem of **optimal transport**:

How to rearrange **f** to look like **g** with the **least amount of “work”**?



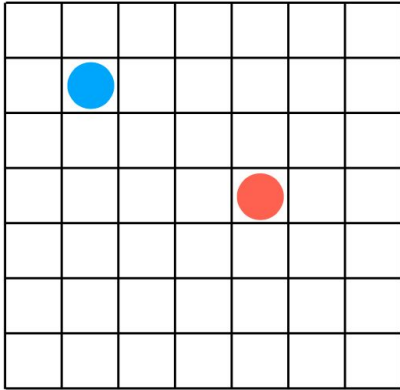
In other words, how can we optimally transport f to g ?

1. Introduction: Why OT as the Metric?

Consider two particles with unit energy.

Image-based approach

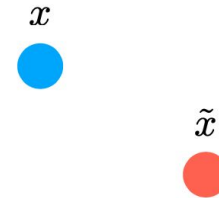
Bin on N-bin grid, represent energy distribution by vectors in \mathbf{R}^N , compute Euclidean distance between vectors



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

regardless of positions

OT-based approach



$$W_p(\mathcal{E}, \tilde{\mathcal{E}}) = \|x - \tilde{x}\|$$

Invaluable if the relative distribution of pixels carries meaning

OT preserves the underlying geometry!

1. Introduction: EMD & Its Generalization

EMD Definition for Jets:

f_{ij} : the amount of energy moved from particle i to particle j .

θ_{ij} : ground distance between particles i and j

$$EMD(E, E') = \min_{f_{ij} \in \Gamma_{E, E'}} \sum_{ij} f_{ij} \theta_{ij}$$

Standard EMD

E : Event E with total energy E
 E' : Event E' with total energy E'
 $E = E'$

Conditions:

$$\Gamma_{E, E'} = \left\{ f_{ij} : f_{ij} \geq 0, \sum_j f_{ij} = E_i, \sum_i f_{ij} = E'_j \right\}$$

E_i : Energy of particle i in event E

E'_j : Energy of particle j in event E'

1. Introduction: EMD & Its Generalization

EMD Definition for Jets:

f_{ij} : the amount of energy moved from particle i to particle j .

θ_{ij} : ground distance between particles i and j

$$EMD(E, E') = \min_{f_{ij} \in \Gamma_{E, E'}} \sum_{ij} f_{ij} \theta_{ij}$$

Standard EMD

E : Event E with total energy E
 E' : Event E' with total energy E'
 $E = E'$

Conditions:

E_i : Energy of particle i in event E
 E'_j : Energy of particle j in event E'

$$\Gamma_{E, E'} = \left\{ f_{ij} : f_{ij} \geq 0, \sum_j f_{ij} = E_i, \sum_i f_{ij} = E'_j \right\}$$

Same as Standard EMD

One Possible Generalization when $E \neq E'$

$$EMD_R(\mathcal{E}, \mathcal{E}') := \min_{f_{ij} \in \tilde{\Gamma}_{\leq(\mathcal{E}, \mathcal{E}')}} \left[\frac{1}{R} \sum_{ij} f_{ij} \theta_{ij} + \left| \sum_i E_i - \sum_j E'_j \right| \right] \quad (1.2)$$

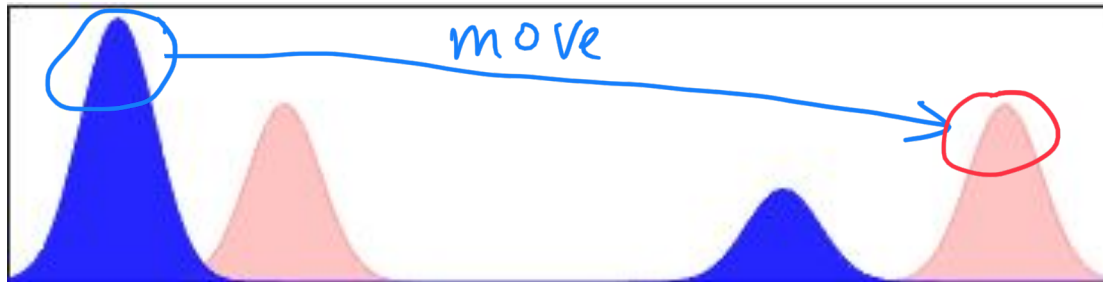
Extra piece to account for the unequal total energy

$$\tilde{\Gamma}_{\leq(\mathcal{E}, \mathcal{E}')} := \left\{ f_{ij} : f_{ij} \geq 0, \sum_j f_{ij} \leq E_i, \sum_i f_{ij} \leq E'_j, \sum_{ij} f_{ij} = \min \left(\sum_i E_i, \sum_j E'_j \right) \right\} \quad (1.3)$$

Different conditions for the unbalanced case

2. Theory of Optimal Transport

Balanced OT: Mass can only be transported, not created or destroyed.
=> Total mass has to be equal.



Unbalanced OT: Mass can be transported, created and destroyed.
=> Total mass can be unequal.



Credit: L. Chizat, G. Peyré, B. Schmitzer, F-X. Vialard
on [G. Peyré's Github](#).

2. Theory of OT: Balanced & Unbalanced

p-Wasserstein Distance

p=1: EMD

p=2: Monge-Kantorovich Distance (W_2); has a (weak) Riemannian structure, thus can be linearized.

$$W_p(\mathcal{E}, \tilde{\mathcal{E}}) = \min_{g_{ij} \in \Gamma(\mathcal{E}, \tilde{\mathcal{E}})} \left(\sum_{ij} g_{ij} \|x_i - \tilde{x}_j\|^p \right)^{1/p}$$

$$\Gamma(\mathcal{E}, \tilde{\mathcal{E}}) = \left\{ g_{ij} : g_{ij} \geq 0, \sum_j g_{ij} = E_i, \sum_i g_{ij} = \tilde{E}_j \right\}$$

2. Theory of OT: Balanced & Unbalanced

p-Wasserstein Distance

p=1: EMD

p=2: Monge-Kantorovich Distance (W_2); has a (weak) Riemannian structure, thus can be linearized.

$$W_p(\mathcal{E}, \tilde{\mathcal{E}}) = \min_{g_{ij} \in \Gamma(\mathcal{E}, \tilde{\mathcal{E}})} \left(\sum_{ij} g_{ij} \|x_i - \tilde{x}_j\|^p \right)^{1/p}$$

$$\Gamma(\mathcal{E}, \tilde{\mathcal{E}}) = \left\{ g_{ij} : g_{ij} \geq 0, \sum_j g_{ij} = E_i, \sum_i g_{ij} = \tilde{E}_j \right\}$$

Hellinger-Kantorovich (HK) Distance:
The unbalanced generalization for W_2 distance; also enjoys a Riemannian structure.

$$\partial_t \rho + \operatorname{div} \omega = \zeta$$

$\zeta=0$: No source => W_2 Distance
 $\zeta \neq 0$: With source => HK Distance

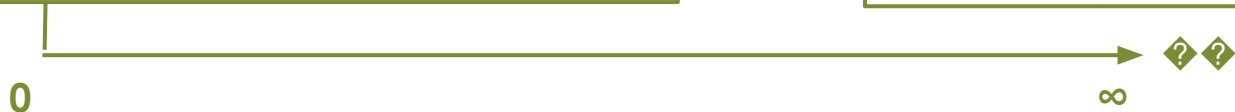
$$J_{\text{HK}, \kappa}(\rho, \omega, \zeta) := \begin{cases} \int_{[0,1] \times \Omega} \left(\left\| \frac{d\omega}{d\rho} \right\|^2 + \frac{\kappa^2}{4} \left(\frac{d\zeta}{d\rho} \right)^2 \right) d\rho & \text{if } \rho \geq 0, \omega, \zeta \ll \rho, \\ +\infty & \text{else.} \end{cases}$$

$$\text{HK}(\mu_0, \mu_1)^2 := \inf \{ J_{\text{HK}}(\rho, \omega, \zeta) \mid (\rho, \omega, \zeta) \in \mathcal{CES}(\mu_0, \mu_1) \}.$$

Intrinsic length scale $\kappa > 0$
controls the relative importance of the transport part of the cost and the creation/destruction part.

$$\text{HK}_\kappa(\mu_0, \mu_1) / \kappa \rightarrow \text{Hellinger distance } (\sim \text{Euclidean})$$

$$\text{HK}_\kappa(\mu_0, \mu_1) \rightarrow W_2(\mu_0, \mu_1)$$



HK Distance: Unnormalized vs. Normalized Measures

REMARK 3.12 (Global mass rescaling behaviour of HK). *Let $\mu_0, \mu_1 \in \mathcal{M}_1(\Omega)$ and $m_0, m_1 \in \mathbb{R}_+$. It was shown in [20, Theorem 3.3] that*

$$(3.14) \quad \text{HK}(m_0 \cdot \mu_0, m_1 \cdot \mu_1)^2 = \sqrt{m_0 \cdot m_1} \cdot \text{HK}(\mu_0, \mu_1)^2 + (\sqrt{m_0} - \sqrt{m_1})^2$$

and if π is optimal in (3.13) for $\text{HK}(\mu_0, \mu_1)^2$ then $\sqrt{m_0 \cdot m_1} \cdot \pi$ is optimal for $\text{HK}(m_0 \cdot \mu_0, m_1 \cdot \mu_1)^2$.

Unbalanced HK on **unnormalized** measures can be obtained from HK from **normalized** measures.

Local mass discrepancies more **important than** the differences in the **total** mass of the measures.

In analysis, **first normalize** all samples before computing HK, then recover the total mass difference via Eq 3.14.

Three Approaches to use OT:

1. **Normalize, then balanced W_2**
2. **Normalize, then unbalanced HK**
3. **Unbalanced HK directly**

2. Theory of OT: Practical Limitation of Exact OT

A dataset with N jets

T_{OT} : Time to compute one pair of OT distance (~ 0.1 secs)

T_2 : Time to compute one pair of Euclidean distance ($\sim 10^{-3}$ secs)

OT for the whole dataset takes time
 $\sim N(N-1)/2 * T_{\text{OT}}$
 \Rightarrow Too long for large datasets!

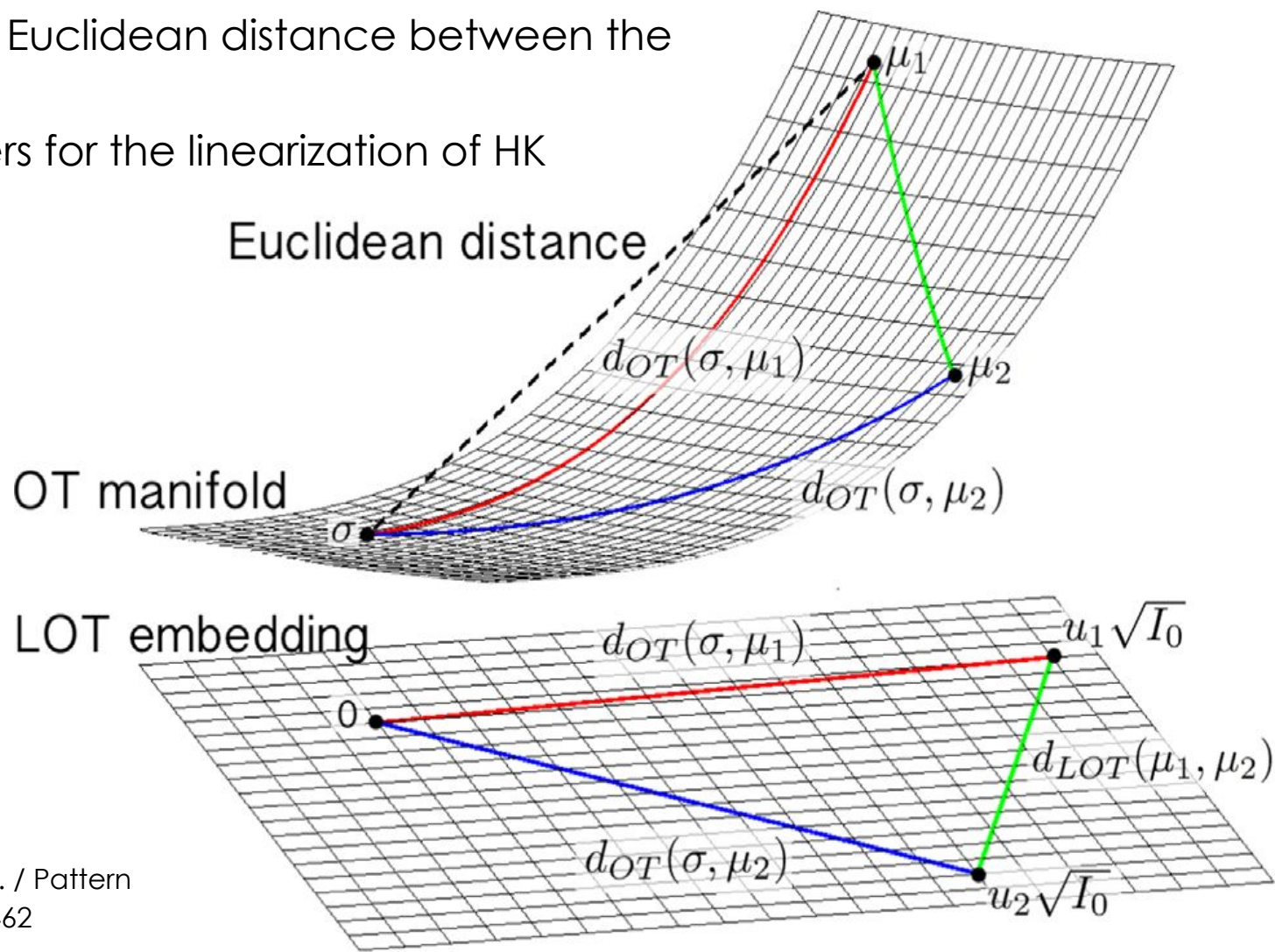
Introduce a **linear version** that only takes time $\sim N * T_{\text{OT}} + N(N-1)/2 * T_2$.

Compute the OT distance between **100k** events takes **~ 16 years** on a desktop.

**Linearized
Optimal Transport!**

2. Theory of OT: Linearized Optimal Transport

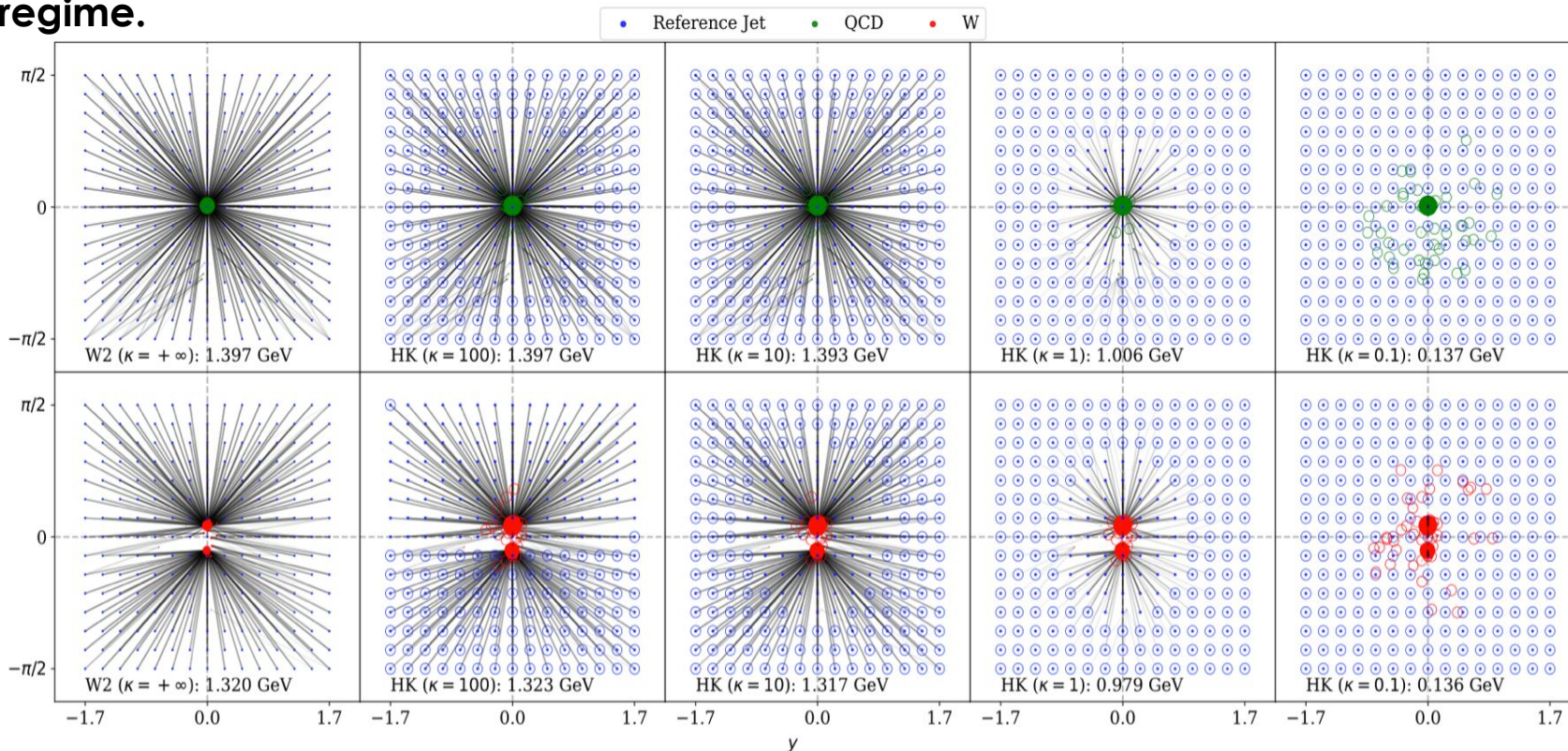
- Project onto 2-Wasserstein tangent plane at a chosen reference event.
- Compute the Euclidean distance between the projections.
- Refer to papers for the linearization of HK distances.



3. LOT for Jet Tagging: OT Plots

Optimal transports between the 15×15 uniform reference measure (blue) and a typical QCD jet (green), or a W jet (red), using W_2 and HK with $\kappa = 100, 10, 1, 0.1$.

The total OT distances between the jets are similar for $\kappa = +\infty, 100, 10$, the transport regime.

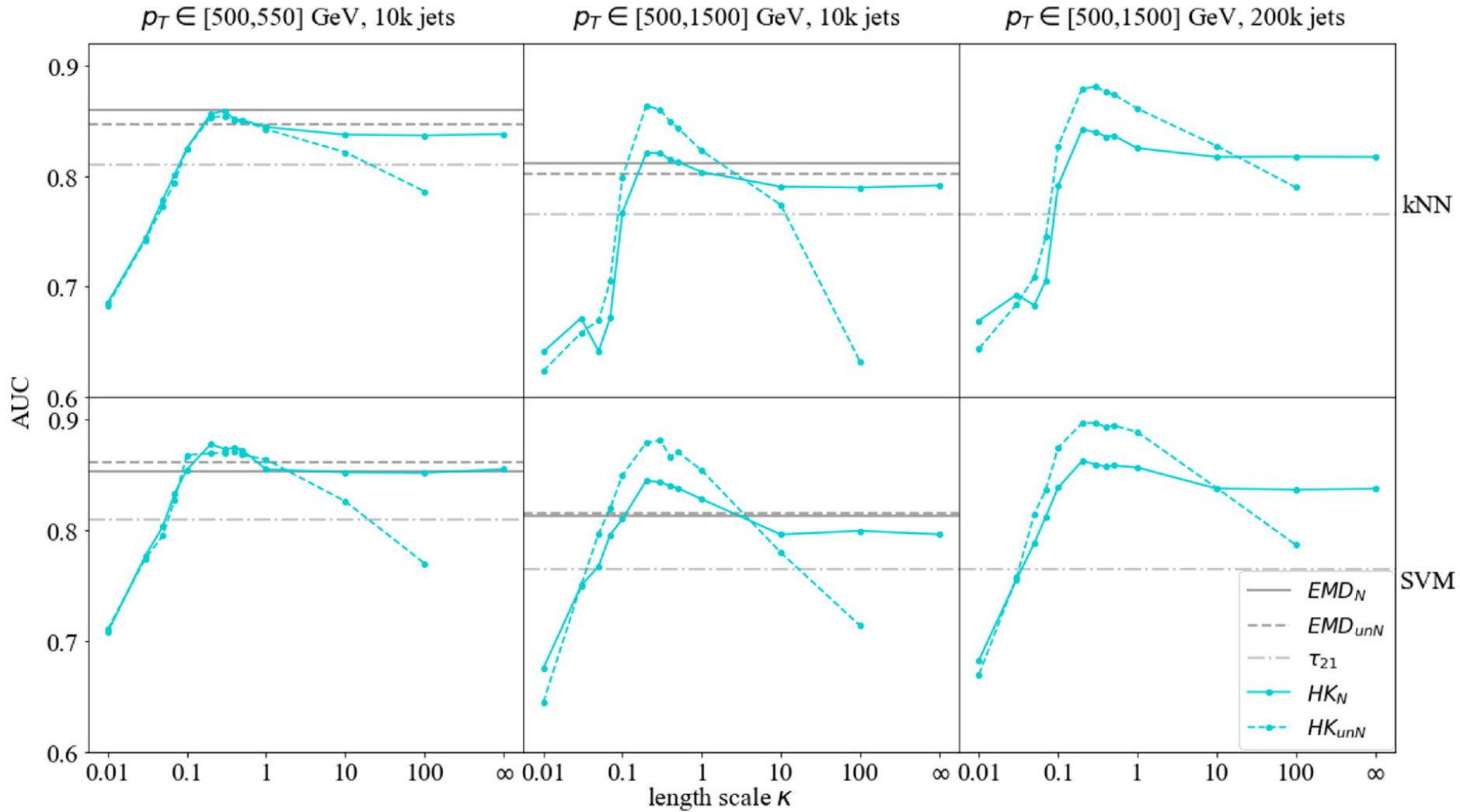


← κ

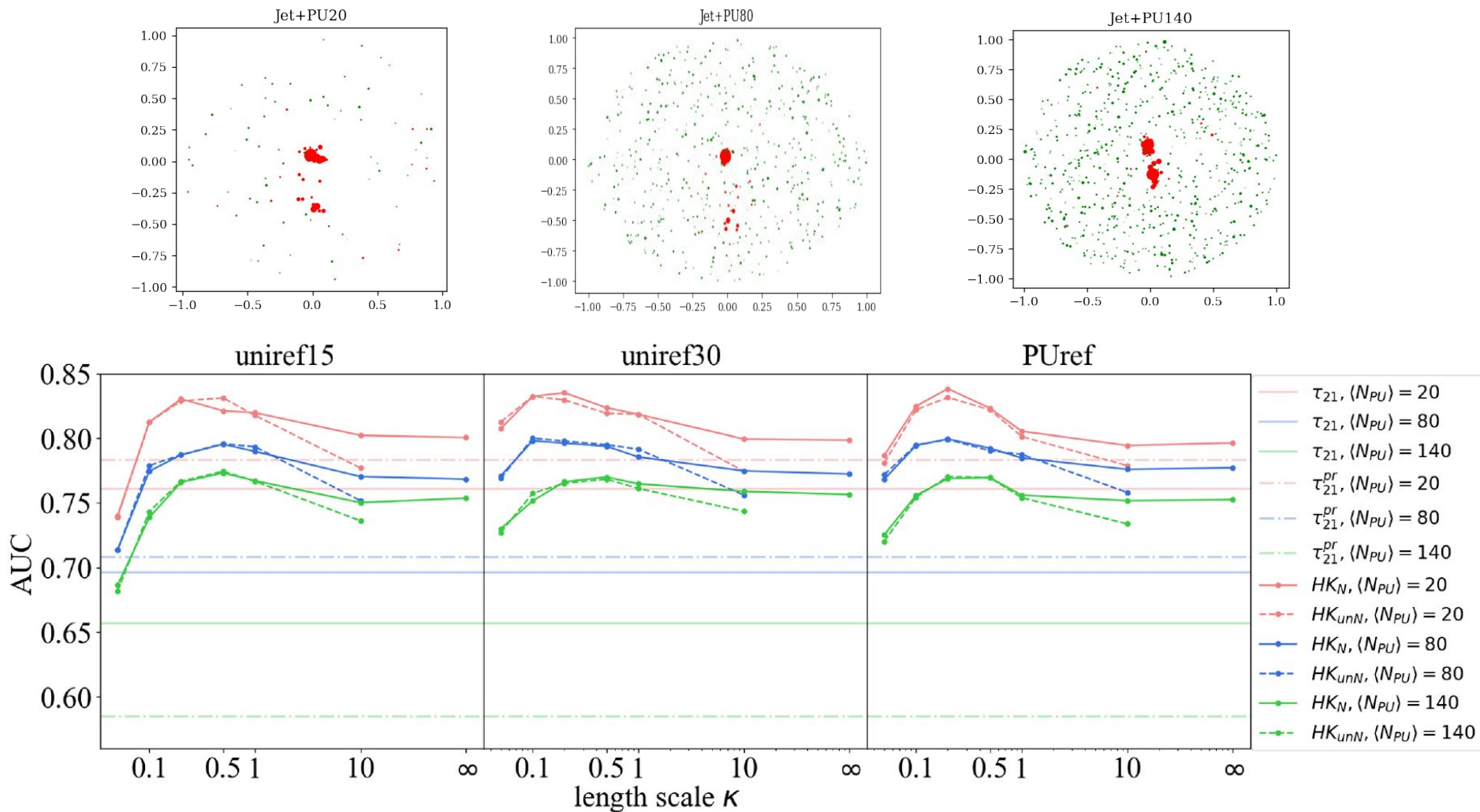
$\kappa = \infty$: original W_2 distance

$\kappa = 0$: image difference

3. LOT for Jet Tagging: W v.s. QCD task using Linearized W_2 and HK distances with various κ



3. Pileup Effect: Is OT relatively insensitive to Pileup?



LOT- W_2 and LOT-HK on 10k WQCD jets with different PUs (Poisson distributions with mean $N_{PU} = 20, 80, 140$) compared with τ_{21} on pruned jets in the same dataset.

4. Summary

- Optimal transport provides a natural metric on the space of collider events with ideal properties.
- Useful for geometrization of LHC data, event classification, unifying description of collider observables...
- ... and now computable on your laptop using Linearized Optimal Transport.



Thank you!

Backup Slides

Balanced OT

p-Wasserstein Distance

p=1: EMD

p=2: Monge-Kantorovich Distance (W_2); has a (weak) Riemannian structure, thus can be linearized.

Kantorovich formulation (static):

$$W_p(\mathcal{E}, \tilde{\mathcal{E}}) = \min_{g_{ij} \in \Gamma(\mathcal{E}, \tilde{\mathcal{E}})} \left(\sum_{ij} g_{ij} \|x_i - \tilde{x}_j\|^p \right)^{1/p}$$

$$\Gamma(\mathcal{E}, \tilde{\mathcal{E}}) = \left\{ g_{ij} : g_{ij} \geq 0, \sum_j g_{ij} = E_i, \sum_i g_{ij} = \tilde{E}_j \right\}$$

Benamou-Brenier formulation (dynamic):

DEFINITION 2.1 (Continuity equation). For $\mu_0, \mu_1 \in \mathcal{M}_1(\Omega)$ we denote by $\mathcal{CE}(\mu_0, \mu_1)$ the set of solutions for the continuity equation on $[0, 1] \times \Omega$, i.e. the set of pairs of measures $(\rho, \omega) \in \mathcal{M}([0, 1] \times \Omega)^{1+d}$ where ρ interpolates between μ_0 and μ_1 and that solve

ρ : charge density
 ω : current density

$$\partial_t \rho + \operatorname{div} \omega = 0$$

No source/sink
Charge conservation

in a distributional sense. More precisely, we require for all $\phi \in C^1([0, 1] \times \Omega)$ that

$$(2.1) \quad \int_{[0,1] \times \Omega} \partial_t \phi \, d\rho + \int_{[0,1] \times \Omega} \nabla \phi \cdot d\omega = \int_{\Omega} \phi(1, \cdot) \, d\mu_1 - \int_{\Omega} \phi(0, \cdot) \, d\mu_0.$$

DEFINITION 2.2 (Wasserstein-2 distance, dynamic formulation [4]). Let $J_W : \mathcal{M}([0, 1] \times \Omega)^{1+d} \rightarrow \mathbb{R} \cup \{\infty\}$ be given by

$$(2.2a) \quad J_W(\rho, \omega) := \begin{cases} \int_{[0,1] \times \Omega} \|\frac{d\omega}{d\rho}\|^2 \, d\rho & \text{if } \rho \geq 0, \omega \ll \rho \\ +\infty & \text{else.} \end{cases}$$

Then for $\mu_0, \mu_1 \in \mathcal{M}_1(\Omega)$ we set

$$(2.2b) \quad W_2(\mu_0, \mu_1)^2 := \inf \{ J_W(\rho, \omega) \mid (\rho, \omega) \in \mathcal{CE}(\mu_0, \mu_1) \}.$$

Unbalanced OT

Hellinger-Kantorovich (HK) Distance:

The unbalanced generalization for W_2 distance; also enjoys a (weak) Riemannian structure, thus can be linearized.

Benamou-Brenier-type formulation (dynamic):

DEFINITION 3.1 (Continuity equation with source). For $\mu_0, \mu_1 \in \mathcal{M}(\Omega)$ we denote by $\mathcal{CES}(\mu_0, \mu_1)$ the set of solutions for the continuity equation with source on $[0, 1] \times \Omega$, i.e. the set of triplets of measures $(\rho, \omega, \zeta) \in \mathcal{M}([0, 1] \times \Omega)^{1+d+1}$ where ρ interpolates between μ_0 and μ_1 and that solve

$$\partial_t \rho + \operatorname{div} \omega = \zeta \quad \text{With source}$$

in a distributional sense. More precisely, we require for all $\phi \in C^1([0, 1] \times \Omega)$ that

$$(3.1) \quad \int_{[0,1] \times \Omega} \partial_t \phi \, d\rho + \int_{[0,1] \times \Omega} \nabla \phi \cdot d\omega + \int_{[0,1] \times \Omega} \phi \, d\zeta = \int_{\Omega} \phi(1, \cdot) \, d\mu_1 - \int_{\Omega} \phi(0, \cdot) \, d\mu_0.$$

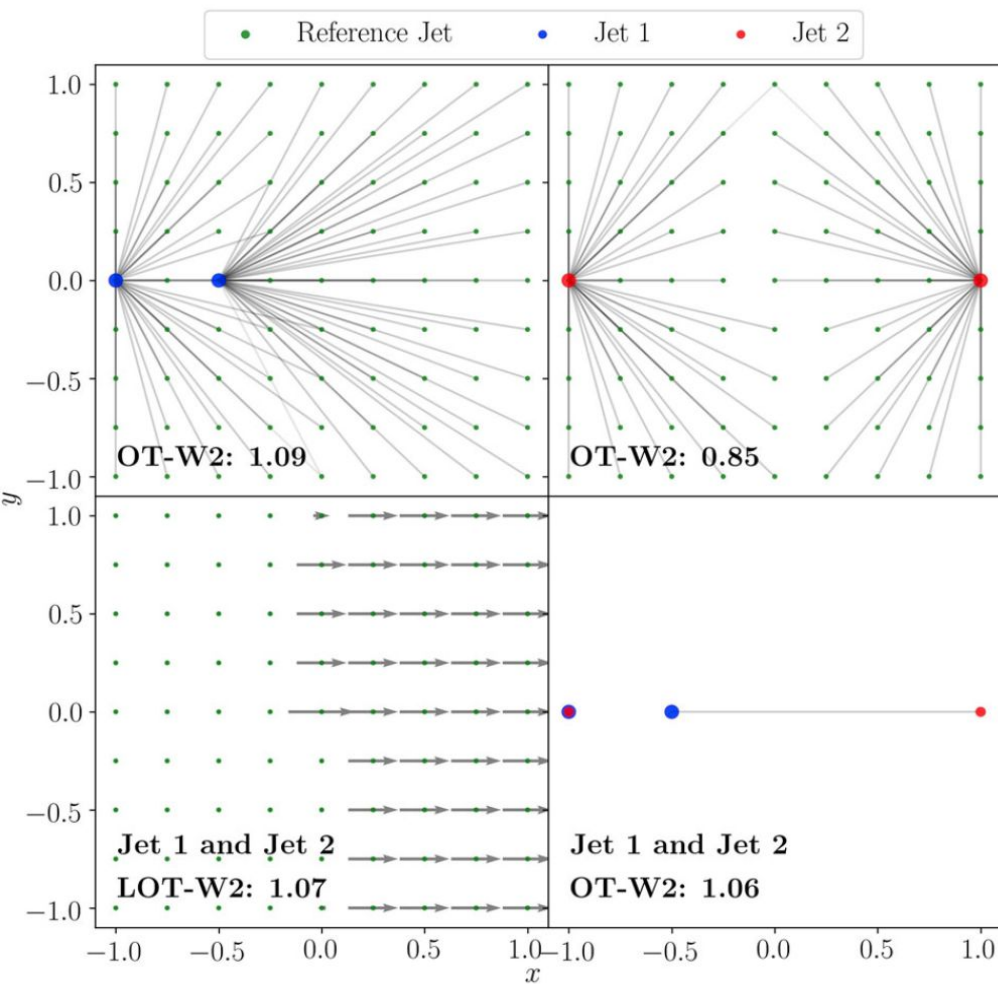
DEFINITION 3.2 (Hellinger–Kantorovich distance, dynamic formulation [19, 9, 22]). Let $J_{\text{HK}} : \mathcal{M}([0, 1] \times \Omega)^{1+d+1} \rightarrow \mathbb{R} \cup \{\infty\}$ be given by

$$(3.2a) \quad J_{\text{HK}}(\rho, \omega, \zeta) := \begin{cases} \int_{[0,1] \times \Omega} \left(\left\| \frac{d\omega}{d\rho} \right\|^2 + \frac{1}{4} \left(\frac{d\zeta}{d\rho} \right)^2 \right) d\rho & \text{if } \rho \geq 0, \omega, \zeta \ll \rho, \\ +\infty & \text{else.} \end{cases}$$

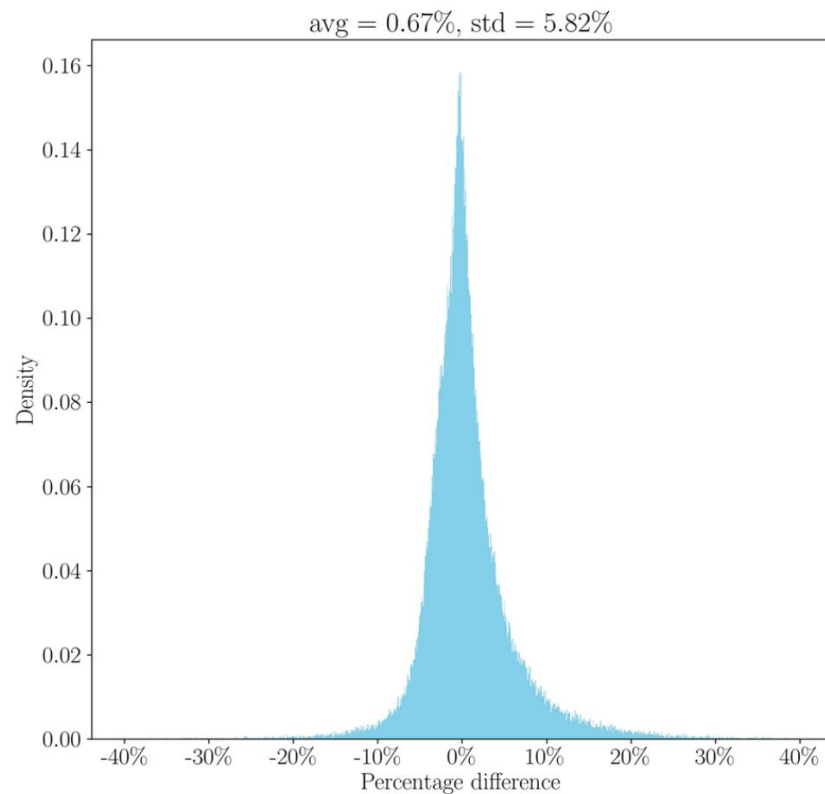
Then for $\mu_0, \mu_1 \in \mathcal{M}_+(\Omega)$ we set

$$(3.2b) \quad \text{HK}(\mu_0, \mu_1)^2 := \inf \{ J_{\text{HK}}(\rho, \omega, \zeta) \mid (\rho, \omega, \zeta) \in \mathcal{CES}(\mu_0, \mu_1) \}.$$

Linearized Optimal Transport for W_2



Difference between LOT- W_2 and W_2 distances for 500 W and QCD jets.



LOT for W_2 : Computation

Ref Jet: $\bar{\sigma} = \sum_{k=1}^{N_\sigma} q_k \delta_{z_k}$

Jet 1: $\mu = \sum_{i=1}^{N_\mu} m_i \delta_{x_i}$

Jet 2: $\nu = \sum_{j=1}^{N_\nu} p_j \delta_{x_j}$

Normalize jet p_T to 1.

Compute OT plans f and g with respect to the ref jet where the **ground metric** is the Euclidean distance squared in the y - ϕ plane.

Compute Barycenters:

$$\bar{x}_k = \frac{1}{q_k} \sum_{i=1}^{N_\mu} f_{k,i} x_i \quad \text{and} \quad \bar{y}_k = \frac{1}{q_k} \sum_{j=1}^{N_\nu} g_{k,j} y_j$$

LOT distance between Jet 1 and 2:

$$d_{aLOT,\sigma}(\mu, \nu)^2 = \min_{\substack{f \in \Pi_{OT}(\sigma, \mu) \\ g \in \Pi_{OT}(\sigma, \nu)}} \sum_{k=1}^{N_\sigma} q_k |\bar{x}_k - \bar{y}_k|^2$$

Linear Coord for Jets: $\mathbf{x}_n = (\sqrt{q_1} a_n^1 \cdots \sqrt{q_{N_\sigma}} a_n^{N_\sigma})^T$

Where a's are the barycenters of the jet.

Our Default Uniform Ref Jet:

225 (15*15) constituent particles

Jet pt = 525 GeV

$|y| \leq 1.7$

$|\phi| \leq \pi/2$

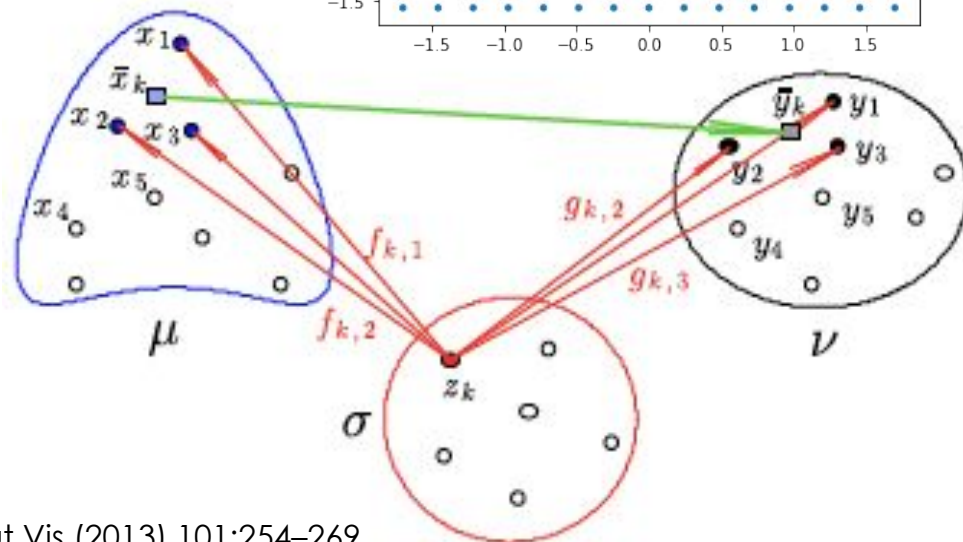
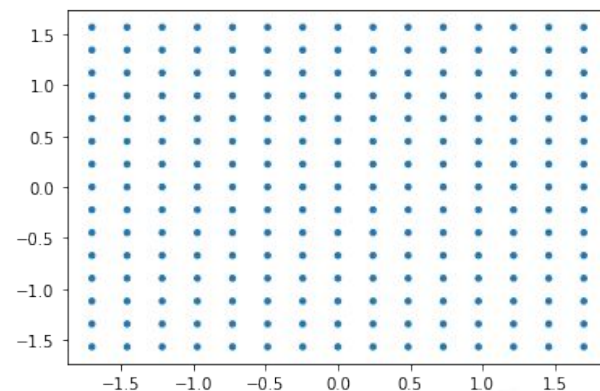


Fig 5, Wei Wang et al. / Int J Comput Vis (2013) 101:254–269

❖ Goal:

- Study the performance of **LOT- W_2** distance paired with various ML models on a number of different jet tagging tasks.
- Study the effect of the length scale **κ in [0.01, 100]** of **LOT-HK** distances on the tagging task of W and QCD discrimination.

❖ Tagging Tasks:

- W v.s. QCD jets (primary)
- t v.s. QCD, t v.s. W
- Higgs v.s. QCD, Higgs v.s. W

$$q\bar{q} \rightarrow Z(\rightarrow \nu\bar{\nu}) + H(\rightarrow b\bar{b})$$

- BSM v.s. QCD, BSM v.s. W

“BSM”: Color sextet scalar

$$q\bar{q} \rightarrow \phi\bar{\phi}$$

$$\phi \rightarrow q\bar{q}$$

$$m_\phi = 100 \text{ GeV}$$

$$\Gamma_\phi = 2 \text{ GeV}$$

❖ Data Generation:

- **MadGraph 2.6.7**: pp collisions at $\sqrt{s} = 14 \text{ TeV}$
- **Pythia 8.243**: Hadronization, multiparton interactions on with default tuning and showering parameters. No detector simulation.
- **FastJet 3.3.2: anti-kt (R=1.0)**. Up to three jets with **p_T in 500-550 GeV** and **$|\mathbf{y}| < 1.7$** are kept.

❖ Jet Preprocessing:

- Centering the jet axis
- Rotation: vertically align the principal component of the constituent p_T in the y-phi plane.

❖ LOT Computation: with a uniform ref jet of 15*15 particles

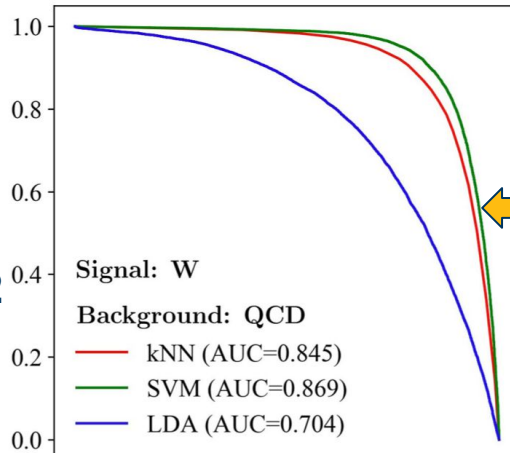
❖ ML Models:

- **LDA**: Supervised Classification & Visualization
- **kNN**: Supervised Classification
 - k in [10, 1000], increment 10
- **SVM**: Supervised Classification
 - C, γ in [10^{-5} , 10^5], increment 10

Full datasets: 140k jets in total for each task (balanced).

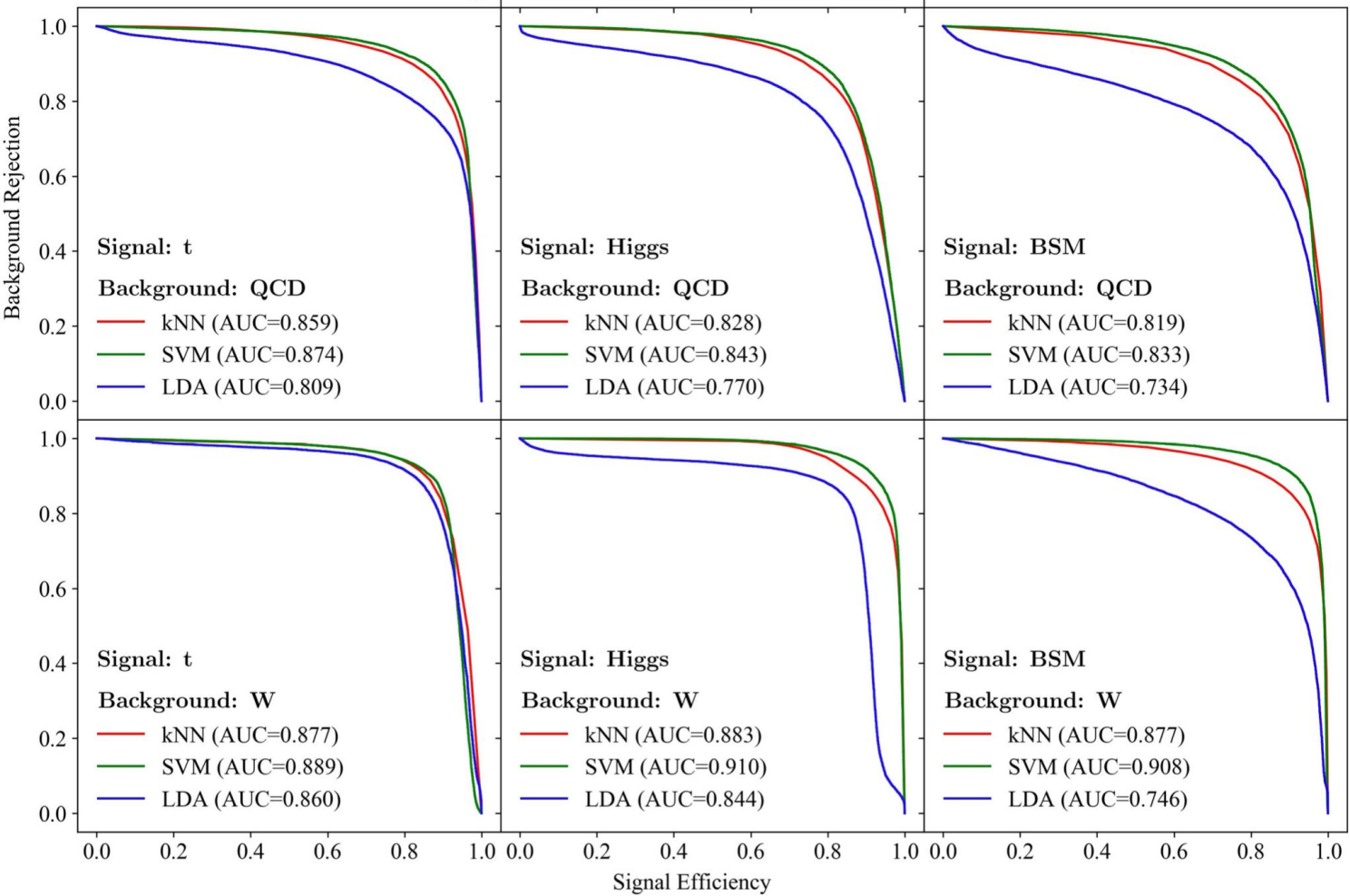
Sample datasets: 10k jets for each task (balanced). Used to pick hyper-parameters of ML models for LOT- W_2 , but as the full datasets for LOT-HK.

OT for Jet Tagging: ML with LOT- W_2 for 7 tasks



Datasets	Model	AUC
Our work	$k=20$ NN-LOT	0.845
	SVM-LOT	0.869
	LDA-LOT	0.704
Komiske, Metodiev, Thaler 1902.02346	$k=32$ NN-EMD	0.887
	$\tau_2^{\beta=1} / \tau_1^{\beta=1}$	0.776
	PFN	0.919
	EFPs	0.917
	EFN	0.904

Comparison with other methods



140k jets whose p_T is in **500-550 GeV**, using **15*15** uniform reference

Jet Tagging: ML with LOT- W_2 for W v.s. QCD

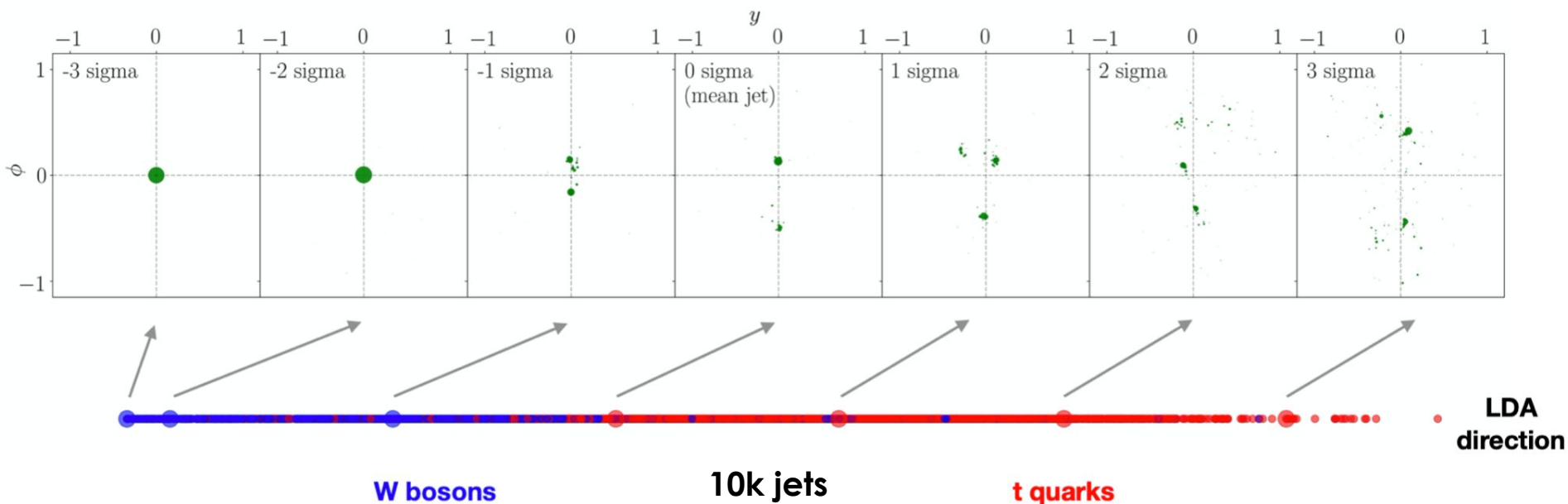
On Sample Dataset (10k jets):

Model	AUC	Best Hyper-param
kNN	0.819	k=20
SVM	0.841	C=1, gamma=100
LDA	0.690	N/A

On Full Dataset (140k jets):

Model	TPR	FPR	Approx. Run Time
kNN	0.803	0.112	4 hours
SVM	0.845	0.108	6 hours
LDA	0.716	0.308	seconds

Jet Tagging: LDA Visualization with LOT- W_2 for t v.s. W



Jet Tagging: ML with LOT-HK of various κ for W v.s. QCD

length scale κ		$+\infty$	100	10	5	1	0.7	0.5	0.3	0.1	0.05	0.01
LDA	AUC	0.694	0.733	0.746	0.747	0.752	0.751	0.748	0.760	0.765	0.763	0.642
	TPR	0.684	0.684	0.703	0.721	0.724	0.740	0.736	0.692	0.704	0.731	0.770
	FPR	0.296	0.218	0.211	0.226	0.220	0.239	0.239	0.171	0.174	0.205	0.486
	run time	several seconds										
kNN [10, 200]	AUC	0.821	0.818	0.819	0.818	0.829	0.841	0.849	0.847	0.821	0.772	0.671
	TPR	0.771	0.763	0.768	0.763	0.760	0.791	0.798	0.809	0.821	0.783	0.733
	FPR	0.128	0.127	0.130	0.126	0.102	0.110	0.100	0.114	0.181	0.238	0.390
	hyperpar. k	30	20	30	20	10	20	10	20	10	10	30
	run time	1.5 hours										
SVM	AUC	0.842	0.842	0.842	0.841	0.849	0.851	0.856	0.853	0.845	0.806	0.694
	TPR	0.817	0.819	0.817	0.819	0.823	0.829	0.832	0.829	0.788	0.741	0.787
	FPR	0.133	0.134	0.134	0.137	0.126	0.127	0.120	0.124	0.099	0.128	0.401
	hyperpar. C	1	1	1	1	1	1	1	1	1	10	10
	hyperpar. γ	100	100	100	100	100	100	100	100	1000	1000	100000
	run time	5 hours										

Results for the WQCD1 (10k jets) dataset using the uniform reference jet

Results for three WQCD datasets using the uniform or the QCD-average reference jet

