ICML
International Conference
On Machine Learning

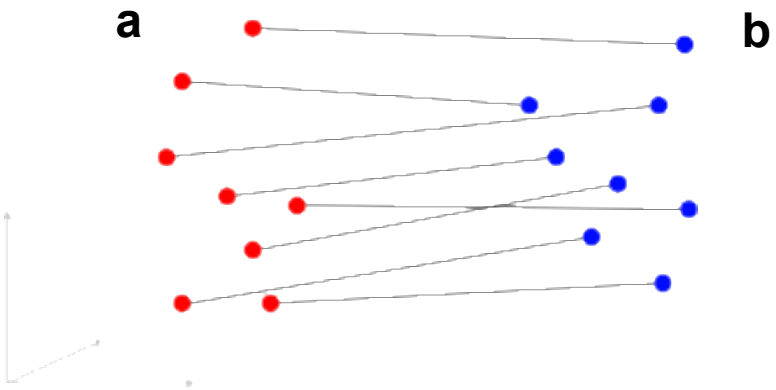
Hierarchical Refinement: Optimal Transport to Infinity and Beyond

Peter Halmos*, Julian Gold*, Xinhao Liu, Benjamin J Raphael

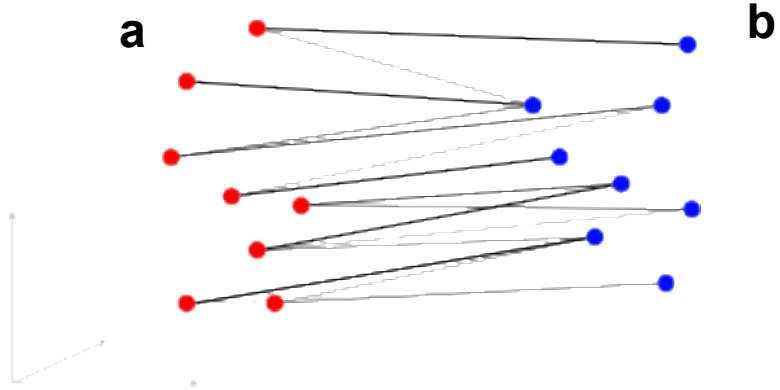
PRINCETON UNIVERSITY

Optimal Transport

- *Optimal transport* finds a least-cost map **T** or coupling **P** between two probability distributions **a** and **b**



Mapping formulation: Function or assignment $x \mapsto y = T(x)$



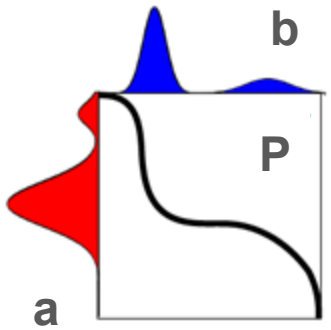
Coupling formulation: Find joint distribution (coupling) P across all pairs (x, y)

Kantorovich Problem

- The primal Kantorovich problem in optimal transport finds the least-cost coupling **P** between **a** and **b** for a distance **C**

$$\min_{\mathbf{P} \in \Pi_{\mathbf{a},\mathbf{b}}} \langle \mathbf{C}, \mathbf{P} \rangle_F$$

- The minimum value of this cost is the *Wasserstein distance* between **a** and **b**



(source: Peyre & Cuturi)

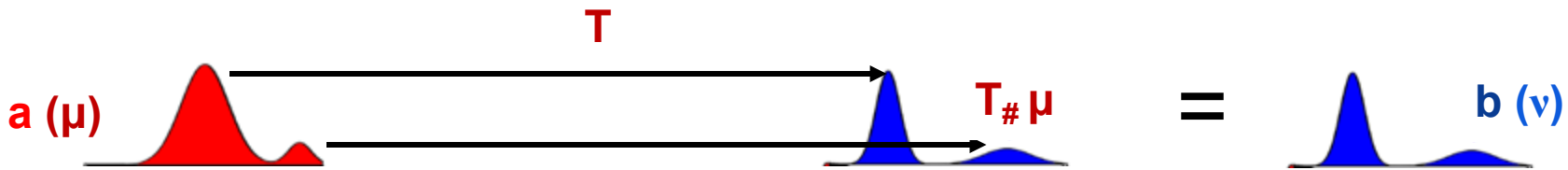
Monge Problem

- Represent **a** and **b** explicitly as measures over their support:

$$\mu = \sum_{i=1}^{n} \mathbf{a}_i \delta_{\mathbf{x}_i} \qquad \nu = \sum_{j=1}^{m} \mathbf{b}_j \delta_{\mathbf{y}_j}$$

- The Monge problem in optimal transport finds the least-cost mapping **T*** for a cost **c**:

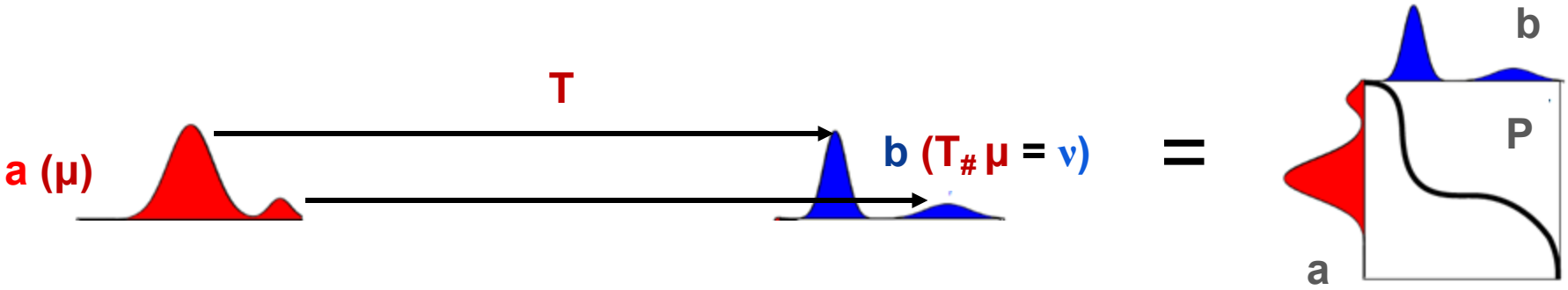
$$T^* = \arg \min_{T_{\#}\mu = \nu} \mathbb{E}_{x \sim \mu}\big[c(x, T(x))\big],$$



*The pushforward (#) is defined for any measurable set $B \subset Y$ by: $T_\sharp \mu(B) := \mu(T^{-1}(B))$

Monge Problem

- When the Monge map exists, the map coincides with the optimal coupling **P** and the problems are equivalent!

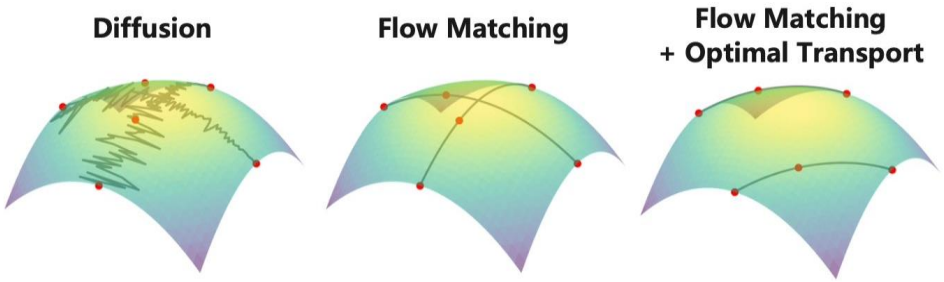


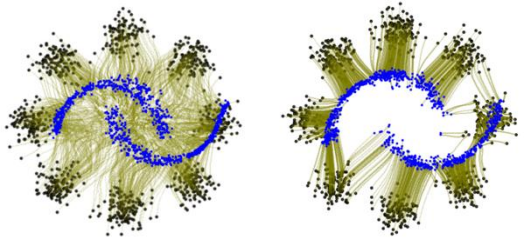
(source: Peyre & Cuturi)

Image Credit: **Korotin ’21, Solomon ’15, Hou ’24, Tong ‘23**

Optimal Transport: Applications

Generative models which obey **least-action** and **minimize energy**

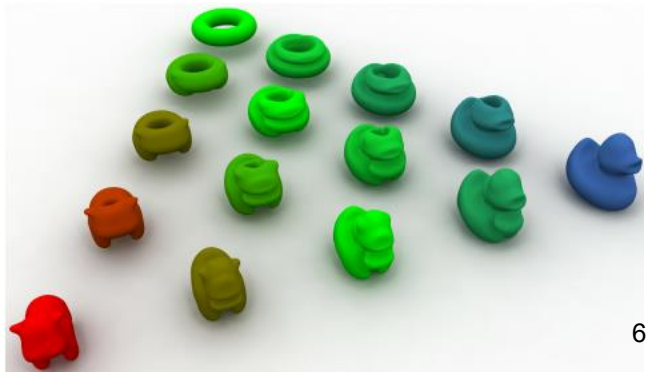




Flow-matching OT Flow-matching

Unpaired Domain-Domain Translation

Optimal registration of **shapes** and **point clouds** in computer vision



Algorithms for Computing OT

- Classical methods (**Tarjan '97, Orlin '97**) solve the Monge problem as a bijective *assignment problem* between *n* points.

$$\text{Time Complexity: } O(n^3)$$
$$\text{Space Complexity: } O(n^2)$$
$$(n \leq 1000, \text{in practice})$$

- The Sinkhorn algorithm (**Cuturi '13**) marked a breakthrough by regularizing the Kantorovich problem with entropy to scale OT in time-complexity

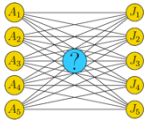
$$\text{Time Complexity: } O(n^2)$$
$$\text{Space Complexity: } O(n^2)$$
$$(n \leq 16000, \text{in practice})$$

$$\min_{\mathbf{P} \in \Pi_{\mathbf{a},\mathbf{b}}} \langle \mathbf{C}, \mathbf{P} \rangle_F - \epsilon \mathrm{H}(\mathbf{P})$$

Current Limitations

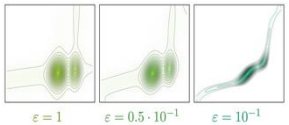
- To use OT on modern datasets [e.g. for OT flow-matching **Lipman ‘22, Tong ’23**; CV & point-cloud registration **Yu ‘21, Qin ’22**; single-cell **Schiebinger ’19**; etc.], one needs to scale to global OT alignments to *millions* of points

Hungarian Algorithm / Network Simplex



O(1k)

Sinkhorn

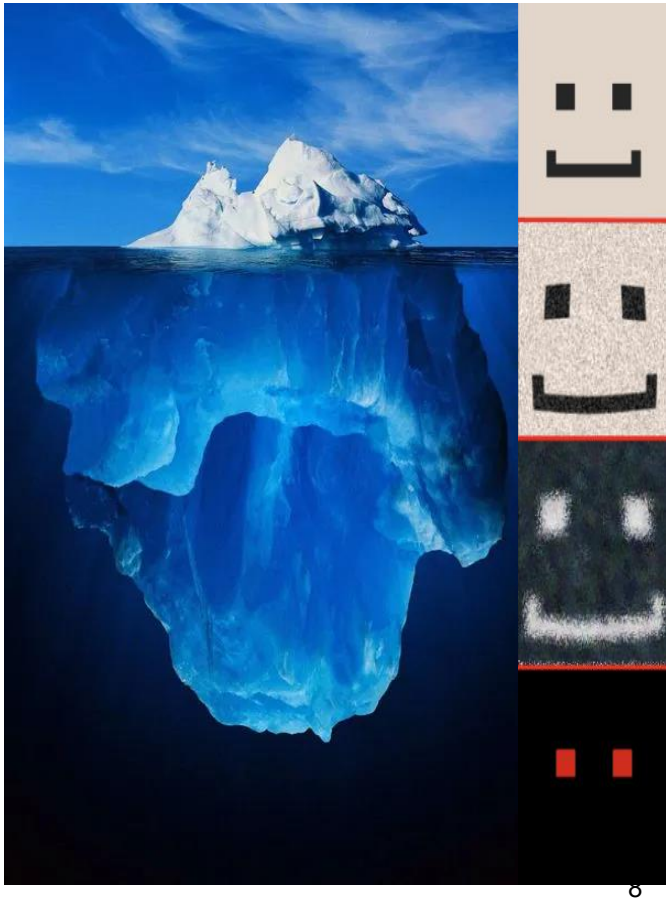


O(10k)

Modern single-cell, brain imaging datasets
O(100k)

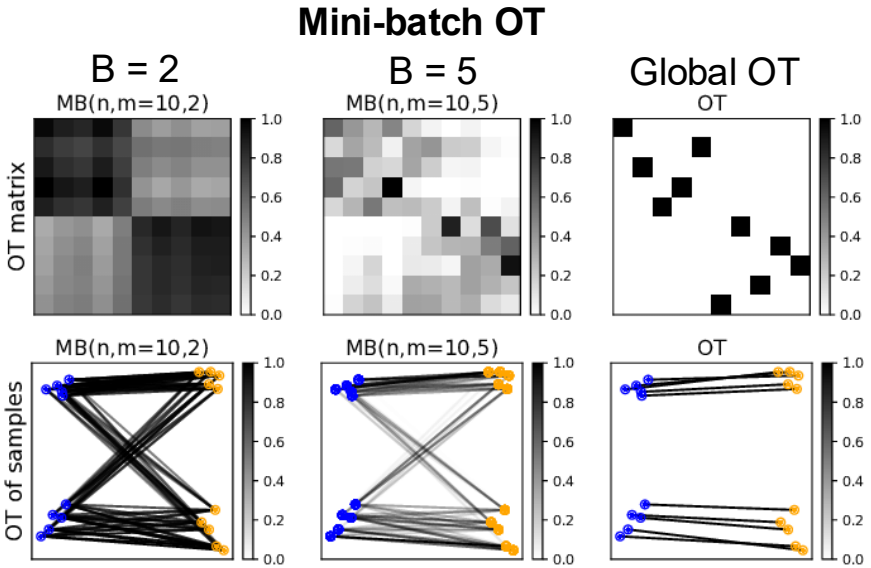
Image classification datasets
O(1-10+million)

Modern text-image datasets
O(100million to +1billion)

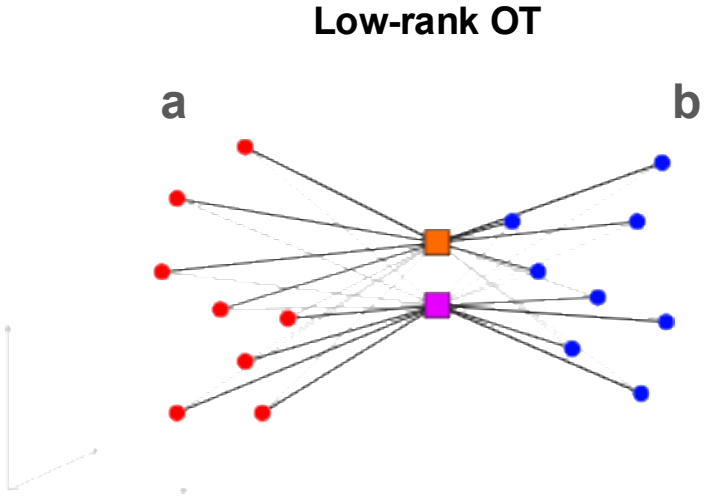


Linear (Time & Space) Approaches to OT

- Mini-batch OT (**Genevay ‘18**) scales linearly, but at the cost of severe batch-biases (**Sommerfeld ‘19; Korotin ‘21; Fatras ‘21**)



- Low-rank OT (**Forrow ’19, Lin ’21, Scetbon ’21, Halmos ’24**) also scales linearly, but forfeits bijective correspondences

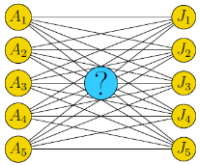


Alignment Quality

Scalability

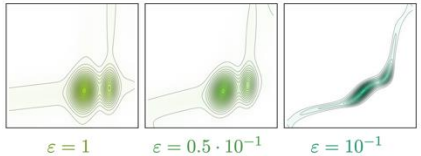
Hungarian Algorithm / Network Simplex

Full-rank and exact,
but O(n^3)

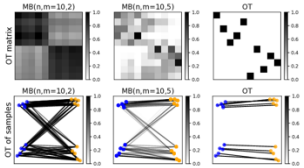


Sinkhorn

Full-rank and approximated
with entropic regularization,
but O(n^2)

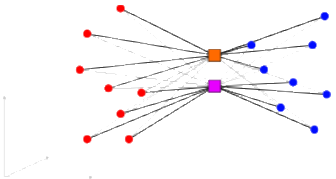


Mini-batch OT



Full-rank, but
alignments are **local**
to each mini-batch

Low-rank OT



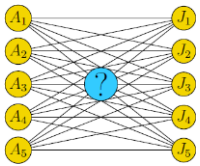
Global but low-
resolution,
constrained by rank

Alignment Quality

Scalability

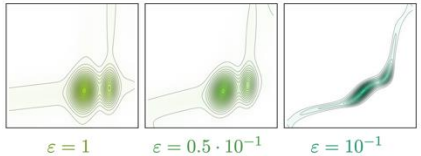
Hungarian Algorithm / Network Simplex

Full-rank and exact, but $O(n^3)$



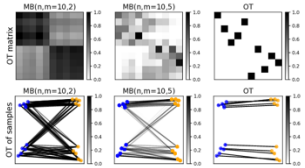
Sinkhorn

Full-rank and approximated with entropic regularization, but $O(n^2)$



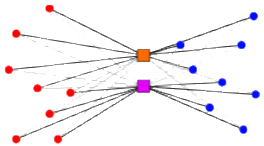
Can we get **global, full-rank** OT with **linear** space/time complexity?

Mini-batch OT



Full-rank, but alignments are **local** to each mini-batch

Low-rank OT

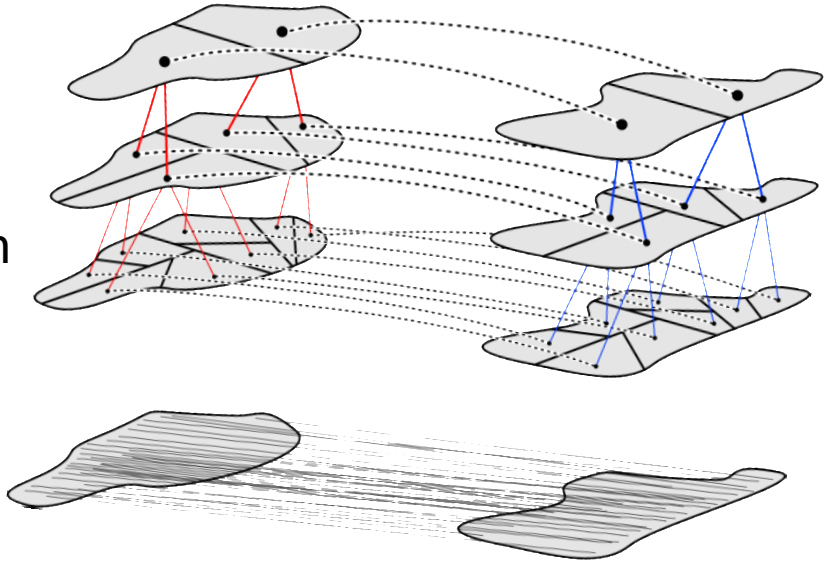


Global but low-resolution, constrained by rank

Hierarchical Refinement: Optimal Transport to Infinity and Beyond

- **Hierarchical Refinement** (HiRef) computes *global, bijective* Monge maps in **linear space** and **log-linear time** on $>10^6$ points.

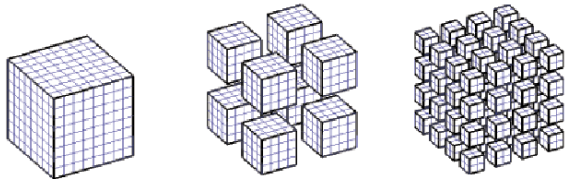
- Hierarchical Refinement is a form of multi-scale OT (**Gerber & Maggioni, '17**), constructing a multi-scale partition of two datasets which culminate in an optimal alignment.



$$T^* = \arg\min_{T_{\#}\mu=\nu} \mathbb{E}_{x\sim\mu}\big[c(x, T(x))\big],$$

Hierarchical Refinement

- The multiscale-OT algorithm of (**Gerber & Maggioni '17**) can find optimal alignments, but it depends on a *pre-defined* grid of the space (akin to dyadic cubes) and suffers a curse of dimensionality



- In HiRef, we prove a key optimality property of the recent technique of low-rank optimal transport (**Forrow '19, Lin '21, Scetbon '21, Halmos '24**) to build these partitions in a manner *intrinsic* to the data

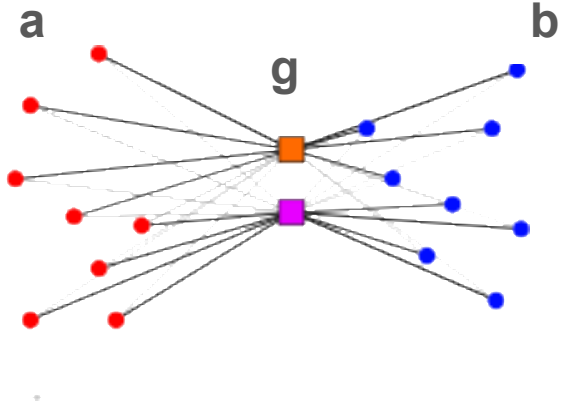
Hierarchical Refinement

- The following optimization problem is a special case of general low-rank optimal transport (**Scetbon '21**) used in Hierarchical Refinement:

$$(\mathbf{Q}^\star, \mathbf{R}^\star) = \operatorname{argmin}_{(\mathbf{Q},\mathbf{R}) \in \Pi_{\mathbf{a},\mathbf{g}} \times \Pi_{\mathbf{b},\mathbf{g}}} \langle \mathbf{C}, \mathbf{Q}\operatorname{diag}(1/\mathbf{g})\mathbf{R}^\top \rangle_F$$

$$\text{s.t.} \quad \mathbf{g} = \mathbf{1}_r/r$$

Low-rank OT



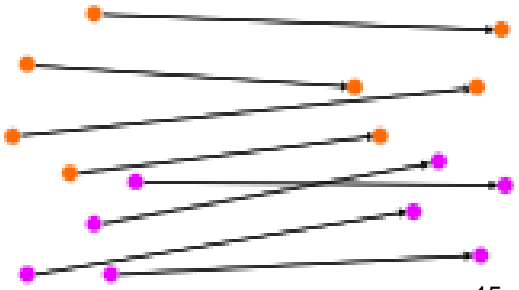
Hierarchical Refinement

- We show that for rank $r = 2$ this problem has optimal solutions corresponding to hard-clustering functions q^*, r^* of the two datasets being aligned (X / $\boldsymbol{\mu}$ and Y / $\boldsymbol{\nu}$)

$$q^\star : X \to [2], \qquad r^\star : Y \to [2]$$

- Using cyclic monotonicity, a property characterizing Monge maps $\mathbf{T}^*$, we prove a key local optimality condition: these clustering functions co-cluster points x with their image $\mathbf{T}^*(x)$ under the optimal mapping

$$q^\star(\mathbf{x}) = r^\star\left(T^\star(\mathbf{x})\right)$$



Points “co-clustered” with their optimal mapping

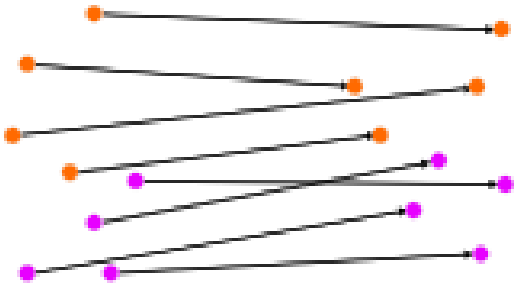
Hierarchical Refinement

- This defines a bipartition of datasets X, Y:

$$X = X_1 \cup X_2, \quad Y = Y_1 \cup Y_2$$

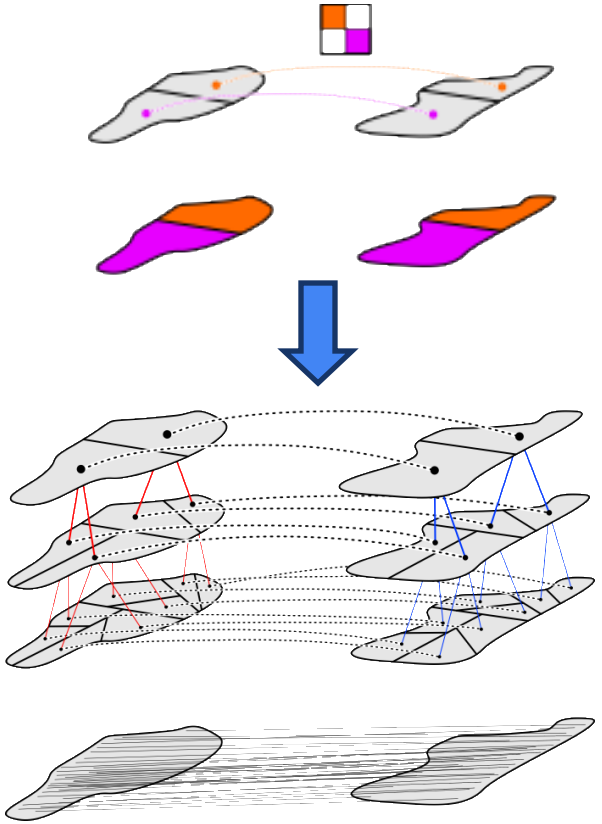
such that the Monge map respects the partition

$$T^\star \big|_{X_1} = Y_1, \quad T^\star \big|_{X_2} = Y_2$$



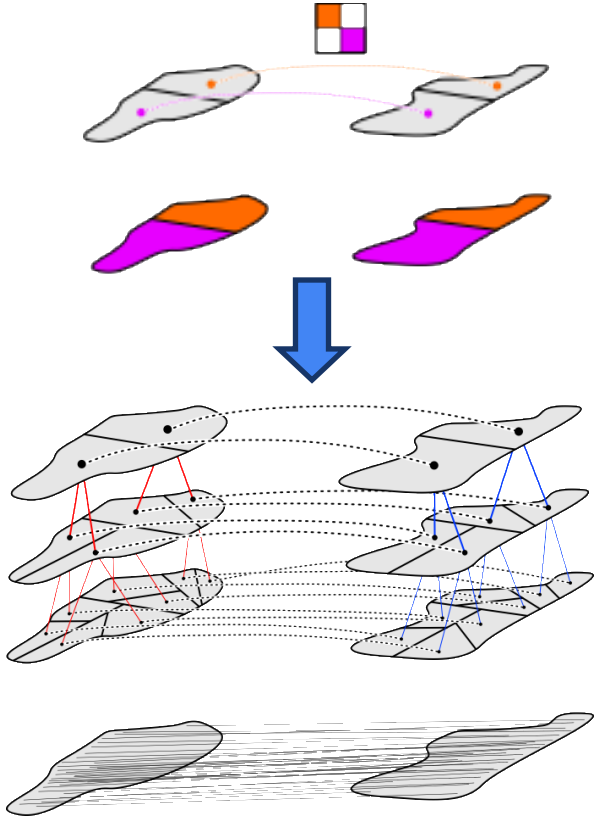
Hierarchical Refinement Algorithm

- We "divide" the points into partitions by solving the low-rank optimization for q^* and r^*
- By maintaining the local $\rightarrow$ global invariant, we can *recurse* on $(X^{(i)}, \mathbf{T}^*(X^{(i)})=Y^{(i)})$
- Recursing to the finest scale recovers a bijective map between point-pairs $(x_i, \mathbf{T}^*(x_i) = y_{\sigma(i)})$



Hierarchical Refinement Algorithm

- This procedure guarantees that we find an *optimal** solution for the Monge map **T** by induction.

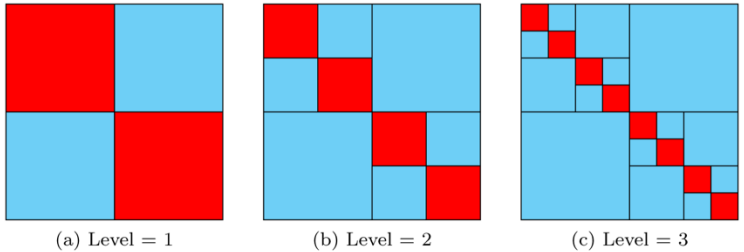


*Conditional on optimality of the low-rank solver

Hierarchical Refinement: Properties

- At each scale t, hierarchical refinement defines an "implicit" hierarchical block coupling:

$$\mathbf{P}^{(t)}_{ij} := \frac{\rho_t}{n^2} \sum_{q=1}^{\rho_t} \delta_{(\mathbf{x}_i, \mathbf{y}_j) \in \Gamma_{t,q}}$$



Hierarchical block-couplings $\mathbf{P}^{(t)}$ across levels. These parallel the structure of Hierarchical Matrices in PDE. **Credit:** Kandappan '23

- Similarly to (**Gerber & Maggioni '17**), we show upper & lower bounds on the implicit iterates of HiRef:

$$0 \leq \langle \mathbf{C}, \mathbf{P}^{(t)} \rangle_F - \langle \mathbf{C}, \mathbf{P}^{(t+1)} \rangle_F \leq \|\nabla c\|_\infty \frac{1}{\rho_t} \sum_{q=1}^{\rho_t} \operatorname{diam}\left(\Gamma_{t,q}\right)$$

Hierarchical Refinement: Complexity

- For n points storing the partitions is $O(n)$ space, and the time-complexity is $O(n\, d \log(n))$ for d the rank of the cost matrix **C**
  (~ ambient dimension d)

- Regarding d as constant (e.g. d ~ 3, 256, 2056 $\ll n$ ~ 1 million, 100 million, 1 billion), Hierarchical Refinement achieves *log-linear* time-scaling

Hierarchical Refinement: Complexity

- When is d small relative to n?
- Costs induced by kernels in an inner-product space such as

$$k(\mathbf{x}_i, \mathbf{x}_i) + k(\mathbf{x}_j, \mathbf{x}_j) - 2k(\mathbf{x}_i, \mathbf{x}_j)$$

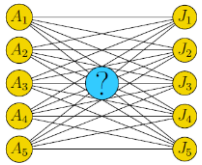
  (most commonly squared Euclidean distance) have distance matrices with rank d ~ ambient dimension of datapoints (see, e.g. **Scetbon '21**)
- Cost matrices not induced by kernels (e.g. Euclidean) satisfying metric properties can still be approximated to rank d in linear time (**Indyk '19**)

Alignment Quality

Scalability

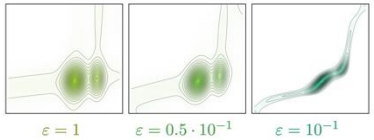
Hungarian Algorithm / Network Simplex

Full-rank and exact, but $O(n^3)$

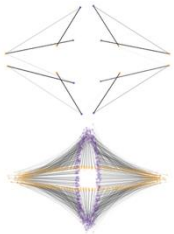


Sinkhorn

Full-rank and approximated with entropic regularization, but $O(n^2)$

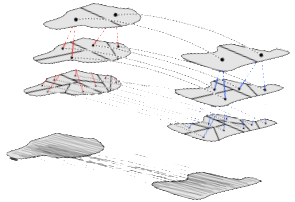


MOP (Multi-Scale OT)



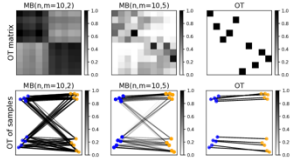
Full-rank and scalable, but solution varies with input grid of space and suffers curse of dimensionality

HiRef



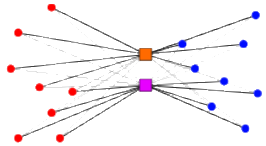
Full-rank, **global**, with O(n) space and O(n log n) time

Mini-batch OT



Full-rank and **scalable**, but alignments are **local** to each mini-batch

Low-rank OT

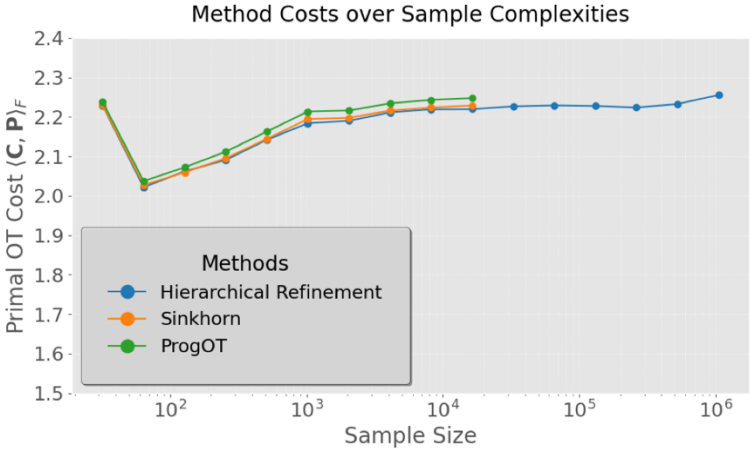


Global but low-resolution, constrained by rank

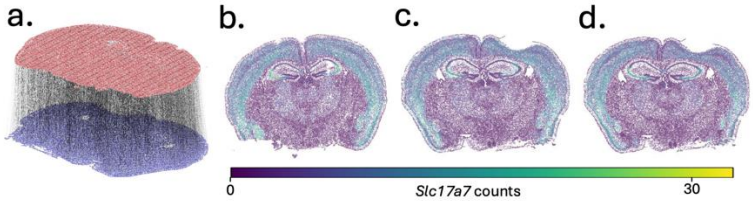
Benchmarking Hierarchical Refinement

- HiRef benchmarked on massive and high-dimensional datasets like
  - Single-cell resolution transcriptomics (n=120k, d=60)
  - ImageNet (n=1.2 million, d=2056)
  - MERFISH Brain Imaging (n=80k, d=2)
  - Synthetic datasets (n=2 million, d=2)

- Solutions comparable to Sinkhorn in cost, and scale beyond 10^6 points (**where most methods do not run!)



Scaling of HiRef on dataset of **Buzun '24**



MERFISH Brain Atlas alignment and predicted gene abundance

Mini-batch in Neural OT Pipelines

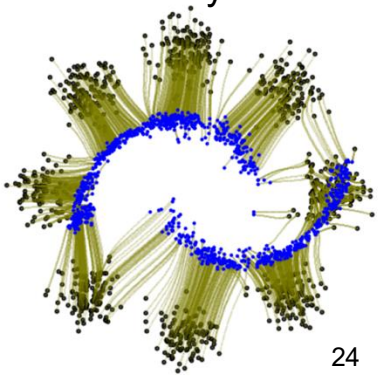
- In generative modeling and deep-learning pipelines which requires OT-driven samples **[Lipman '22, Tong '23, Seguy '18, Yu '21, Qin '22, etc]** for some loss ℓ and network f_θ the paradigm for scaling is to use mini-batch OT

For every training iteration t

$$\hat{\mu}_t \sim \text{Sample}_B(\mu), \quad \hat{\nu}_t \sim \text{Sample}_B(\nu)$$ ← Coupling only local to 2 independent size-B batches

$$P_t^\varepsilon = \arg \min_{P \in \Pi(\hat{\mu}_t, \hat{\nu}_t)} \langle C, P \rangle - \varepsilon H(P),$$ ← Entropic bias, online $O(B^2)$ OT-solve every iteration!

$$\mathcal{L}_{\text{MB}}(\theta) = \mathbb{E}_{(x,y)\sim P_t^\varepsilon}\left[\ell(f_\theta(x), y)\right]$$



OT-CFM (**Tong et al '23**)

HiRef application to Neural OT Pipelines

- With Hierarchical Refinement we can instead precompute pairs globally once and for all, and treat them as supervised training pairs $\left(x_i, \boldsymbol{T}(x_i) = y_{\sigma(i)}\right)_{i=1}^{n}$

$$T^* = \arg\min_{T_{\#}\mu=\nu} \mathbb{E}_{x\sim\mu}\left[c(x, T(x))\right], \quad \leftarrow \text{Compute global Monge map offline in } O(n\log(n))$$

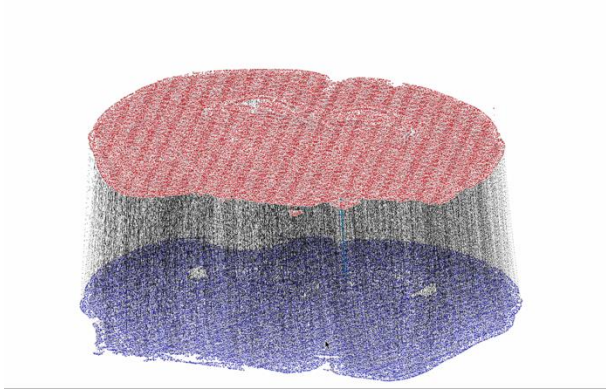
$$\mathcal{L}_{\text{HiRef}}(\theta) = \mathbb{E}_{x\sim\mu}\left[\ell(f_\theta(x), T^*(x))\right] \quad \leftarrow \text{No cost, no bias! Just } O(1) \text{ index into precomputed global pairs}$$

Summary: Hierarchical Refinement (HiRef)

- A multi-scale optimal transport method which scales optimal transport with *linear* complexity and computes bijective mappings between millions of points

- Uses a key local → global optimality property of low-rank OT to guarantee an optimal solution with a divide-and-conquer approach

- Opens the door to applications and datasets previously infeasible for optimal transport



Alignment on MERFISH Brain Imaging (**Clifton et al '23**)

Thank you!

Code: https://github.com/raphael-group/HiRef

Acknowledgments

Raphael Group

Prof. Ben Raphael
Dr. Julian Gold
Dr. Hirak Sarkar
Dr. Yihang Shen
Dr. Mike Wilson
Dr. Hongyu Zheng
Viola Chen
Gillian Chu
Peter Halmos
William Howard-Snyder
Gary Hu
Akhil Jakatdar
Xinhao Liu
Sereno Lopez-Darwin
Henri Schmidt
Ahmed Shuaibi
Richard Zhang
Clover Zheng

PRINCETON UNIVERSITY

NIH NATIONAL CANCER INSTITUTE Informatics Technology for Cancer Research

SCHMIDT FUTURES