Transport Clustering: Solving Low-Rank Optimal Transport via Clustering

Henri Schmidt,* Peter Halmos,* Benjamin J Raphael (*Equal contribution)

ICML International Conference On Machine Learning

PRINCETON UNIVERSITY

Paper:

Background

Optimal transport (OT) finds a least-cost coupling P between any two probability measures μ and ν. In discrete OT the data X and Y are supported on two discrete measures with weights $a \in \Delta_n$, $b \in \Delta_m$:

$$\mu = \sum_{i=1}^{n} \boldsymbol{a}_i \delta_{x_i}, \qquad \nu = \sum_{j=1}^{m} \boldsymbol{b}_j \delta_{y_j},$$

For a given notion of cost $c(\cdot,\cdot)$ one solves the following minimization for the optimal coupling:

$$\min_{\mathbf{P} \in \Pi(\boldsymbol{a},\boldsymbol{b})} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{P}_{ij}\, c(x_i, y_j) = \min_{\mathbf{P} \in \Pi_{\boldsymbol{a},\boldsymbol{b}}} \langle \mathbf{C}, \mathbf{P} \rangle_F$$

Over the Transportation Polytope

$\Pi(a,b) = \{P \in \mathbb{R}_+^{n \times m} : P 1_m = a,\ P^\top 1_n = b\}$

Low-Rank Optimal Transport (Forrow '19, Scetbon '21, Lin '21, Halmos '24) factors the coupling P into low-rank sub-couplings with a non-negative rank constraint. Following Cohen and Rothblum '93 and Scetbon '21, Low-Rank OT can be formulated as following optimization problem.

$$\min_{\substack{\mathbf{Q} \in \Pi(\boldsymbol{a},\boldsymbol{g}), \mathbf{R} \in \Pi(\boldsymbol{b},\boldsymbol{g}) \\ \boldsymbol{g} \in \Delta_K}} \langle \mathbf{C}, \mathbf{Q}\, \mathrm{diag}(\boldsymbol{g}^{-1}) \mathbf{R}^\top \rangle_F$$

K-Means seeks a partition π = {C_k} of X minimizing an energetic cluster distortion. This can be written in a *mean-free* form, which corresponds to a specialized cost for Generalized K-Means (Scetbon '22):

$$\min_{\mathbf{Q} \in \Pi_{\bullet}(\boldsymbol{u}_n, \cdot)} \langle \mathbf{C}, \mathbf{Q}\, \mathrm{diag}(1/\mathbf{Q}^\top \mathbf{1}_n)\, \mathbf{Q}^\top \rangle_F$$

Why Low-Rank Optimal Transport?

- Can avoid the $\mathcal{O}(n^{-2/d})$ curse of full-rank OT in convergence to EMD of underlying measures $\hat{W}_2^2(\mu_n, \nu_n) \to W_2^2(\mu, \nu)$
- Generalizes K-Means to Co-Clustering across >1 dataset
- Low-Rank OT is the only way to find structured couplings. Full-rank OT is *unstructured* with a flat singular spectrum.

The Gap:

- Current algorithms use local heuristics with no theoretical guarantees beyond convergence to local minima
- Non-convex, NP-hard problem similar to NMF (Lee & S. '99)

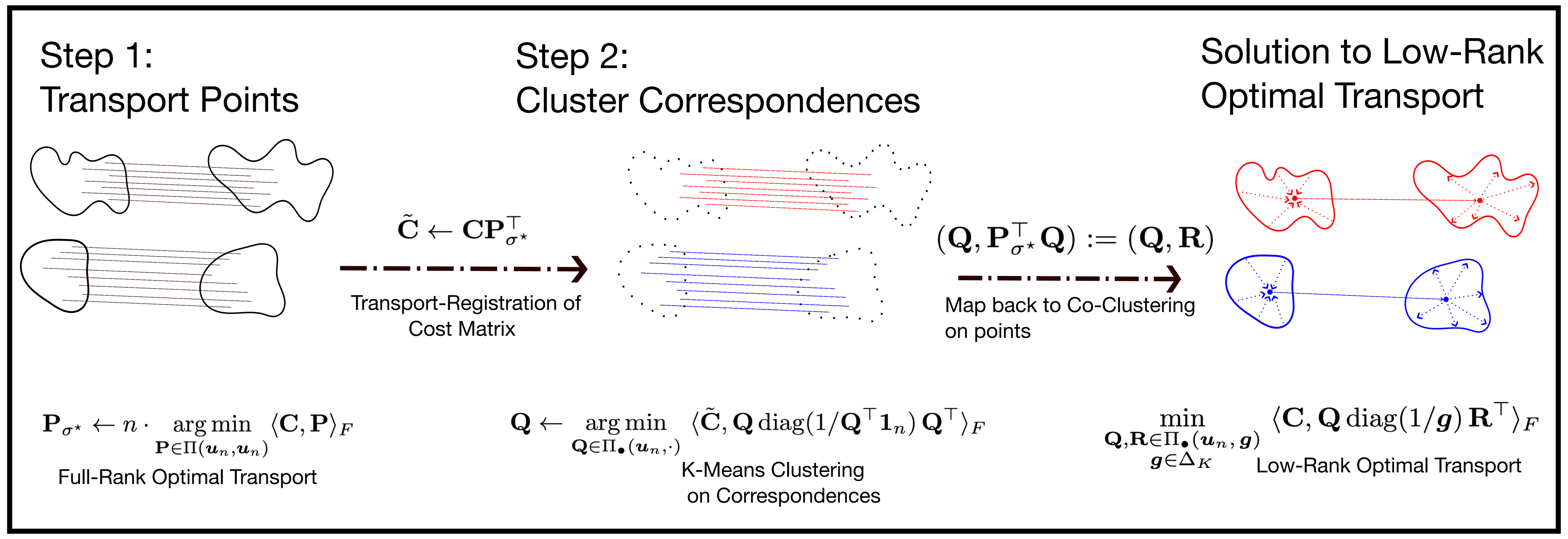


Illustration of Transport Clustering

Transport Clustering

Low-Rank OT can be re-written as an optimization over a clustering matrix **Q** and a permutation $\mathbf{P}_\sigma$

$$\min_{\substack{\mathbf{Q} \in \Pi_{\bullet}(\boldsymbol{u}_n, \cdot), \\ \mathbf{P}_\sigma \in \mathcal{P}_n}} \langle \mathbf{C}\mathbf{P}_\sigma^\top, \mathbf{Q}\, \mathrm{diag}(1/\mathbf{Q}^\top \mathbf{1}_n)\, \mathbf{Q}^\top \rangle_F,$$

Our main question is then:

Is there an efficiently computable choice of permutation matrix $\mathbf{P}_\sigma$ that accurately reduces low-rank optimal transport to the generalized K-means problem?

We answer this in the affirmative. We show given an algorithm for Generalized K-Means that taking the full-rank optimal transport permutation P_{σ^*} — the solution to a *convex problem* — yields a constant-factor approximation algorithm Transport Clustering or TC (Algorithm 1) for Low-Rank Optimal Transport.

Theoretical Results

Theorem 4.1 (Upper-Bounds): Let $C \in \mathbb{R}^{n \times n}$ be a cost induced by i) a metric of negative type, ii) a kernel cost, or iii) a cost satisfying the triangle inequality. If P_σ* denotes the full-rank optimal transport plan for C and $\tilde{C} = CP_{\sigma^*}^\top$ is the transport-registered cost, then the solution to Algorithm 1 satisfies

$$\min_{\mathbf{Q} \in \Pi_{\bullet}(\boldsymbol{u}_n, \cdot)} \langle \tilde{\mathbf{C}}, \mathbf{Q}\, \mathrm{diag}(1/\mathbf{Q}^\top \mathbf{1}_n)\, \mathbf{Q}^\top \rangle_F \leq \begin{cases} (1+\gamma) \cdot \mathrm{OPT} & \textit{if } \mathbf{C} \textit{ is of negative type}, \\ (1+\gamma+\sqrt{2\gamma}) \cdot \mathrm{OPT} & \textit{if } \mathbf{C} \textit{ is a kernel cost}, \\ (1+\gamma+\rho) \cdot \mathrm{OPT} & \textit{if } \mathbf{C} \textit{ is a metric}, \end{cases}$$

OPT = Optimal Low-Rank OT cost.

$\gamma \in [0,1]$ = Approximation Ratio of full to low rank solutions $\mathrm{OPT}(n)/\mathrm{OPT}(K)$.

$\rho \in [0,1]$ = Asymmetry Coefficient on cluster variances.

Theorem 4.3 If we are given black-box $(1+\epsilon)$ approximation algorithms for K-Means and K-Medians, then for (Q^*, R^*) the solution to TC with initialized transport-registration satisfies

$$(1+\epsilon)^{-1} \cdot \langle \mathbf{C}, \mathbf{Q}^* \,\mathrm{diag}(1/(\mathbf{Q}^*)^\top \mathbf{1}_n)\, (\mathbf{R}^*)^\top \rangle_F \leq \begin{cases} (2+2\gamma) \cdot \mathrm{OPT} & \textit{if } \mathbf{C} \textit{ is of negative type}, \\ (1+\gamma+\sqrt{2\gamma}) \cdot \mathrm{OPT} & \textit{if } \mathbf{C} \textit{ is a kernel cost}, \\ (2+2\gamma+2\rho) \cdot \mathrm{OPT} & \textit{if } \mathbf{C} \textit{ is a general metric}, \end{cases}$$

*In **Prop. 4.2 (Lower-Bounds for Algorithm 1)** we find hard instances proving the bounds are tight.

Transport Clustering Algorithm

Algorithm 1 Transport Clustering (TC)

1: **Input:** Cost matrix $\mathbf{C}$ and rank K.

2: **Step 1 (Transport):** Compute the optimal full-rank plan $\mathbf{P}_{\sigma^*}$:

$$\mathbf{P}_{\sigma^*} \leftarrow n \cdot \arg\min_{\mathbf{P} \in \Pi(\boldsymbol{u}_n, \boldsymbol{u}_n)} \langle \mathbf{C}, \mathbf{P} \rangle_F$$

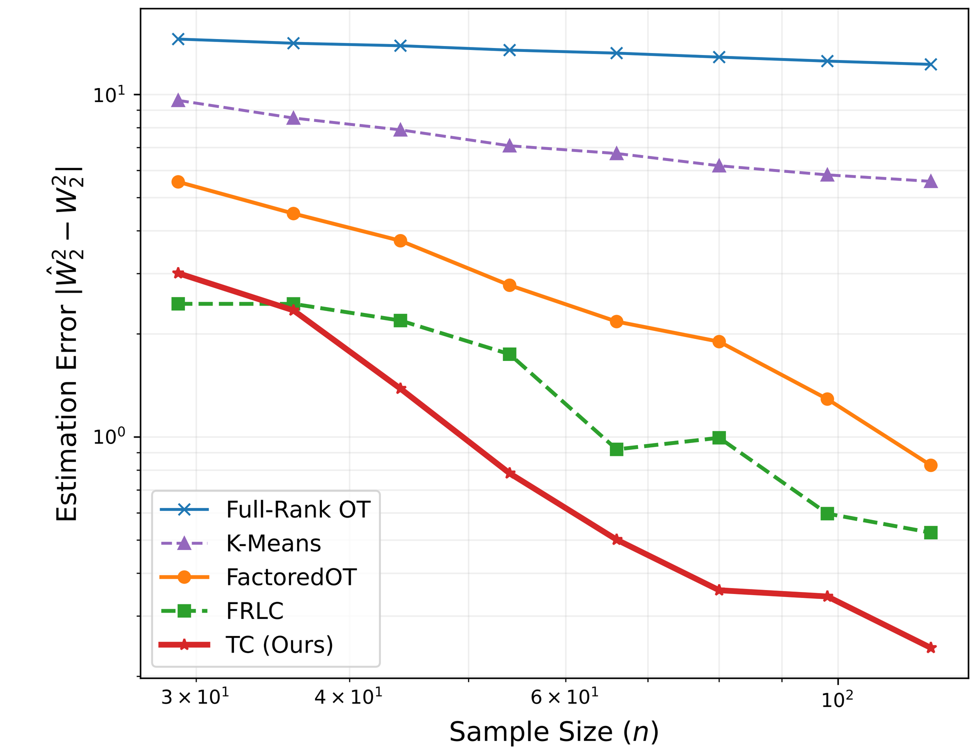
3: **Step 2 (Clustering):** Register the cost $\tilde{\mathbf{C}} \leftarrow \mathbf{C}\mathbf{P}_{\sigma^*}^\top$ and solve generalized K-means for $\mathbf{Q}$:

$$\mathbf{Q} \leftarrow \arg\min_{\mathbf{Q} \in \Pi_{\bullet}(\boldsymbol{u}_n, \cdot)} \langle \tilde{\mathbf{C}}, \mathbf{Q}\, \mathrm{diag}(1/\mathbf{Q}^\top \mathbf{1}_n)\, \mathbf{Q}^\top \rangle_F$$

4: **Output:** The low-rank factors $(\mathbf{Q}, \mathbf{P}_{\sigma^*}^\top \mathbf{Q})$.

Experimental Evaluation

Using the Statistical EMD Estimator of Forrow '19 for Low-Rank couplings, assessed convergence of EMD to analytic truth $\hat{W}_2^2(\hat{\mu}_n, \hat{\nu}_n) \to W_2^2(\mu, \nu)$ in (1) Bures-Wasserstein, and (2) Fractured Hypercube benchmark (Forrow '19).



Improved performance over local solvers for Low-Rank OT such as FRLC (Halmos '24), LOT (Scetbon '21), Factored-OT (Forrow '19), and LatentOT (Lin '21)

Table 1. Comparison of low-rank OT methods across three datasets: CIFAR-10 (n = 60,000), smallest mouse embryo split (n = 18,819), and largest mouse embryo split (n = 131,040).

Dataset	Method	Rank	OT Cost ↓	AMI (A/B) ↑	ARI (A/B) ↑	CTA ↑
CIFAR-10 (60,000)	TC	10	**231.200**	**0.478 / 0.476**	**0.358 / 0.356**	**0.412**
	FRLC	10	235.950	0.411 / 0.407	0.281 / 0.277	0.351
	LOT	10	234.733	0.430 / 0.427	0.306 / 0.303	0.358
Mouse embryo E8.5 → E8.75 (18,819)	TC	43	**0.506**	**0.639 / 0.617**	**0.329 / 0.307**	**0.722**
	FRLC	43	0.553	0.556 / 0.531	0.217 / 0.199	0.525
	LOT	43	0.520	0.605 / 0.592	0.283 / 0.272	0.611
Mouse embryo E9.5 → E9.75 (131,040)	TC	80	**0.389**	**0.554 / 0.551**	**0.172 / 0.169**	**0.564**
	FRLC	80	0.399	0.491 / 0.487	0.116 / 0.115	0.447
	LOT	80	–	– / –	– / –	–