TOWARDS OPTIMAL TRANSPORT WITH GLOBAL INVARIANCES

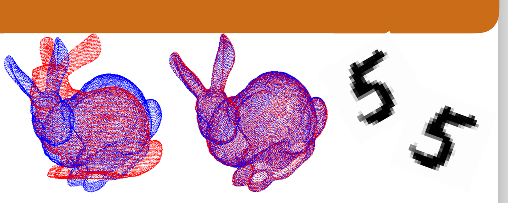
David Alvarez-Melis, Stefanie Jegelka & Tommi S. Jaakkola | **MIT CSAIL**

Summary

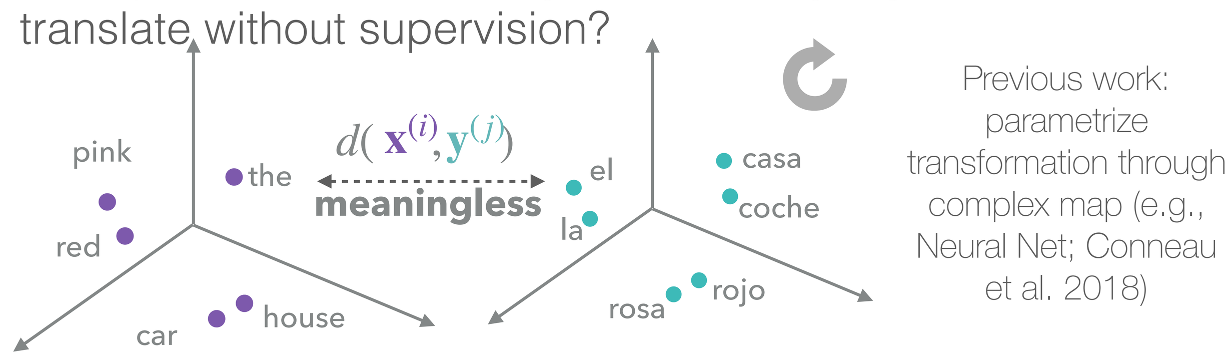
- A generalization of the discrete optimal transport problem that allows for global geometric invariances to be encoded
- Solved through alternating minimization. Under simple conditions, problem simplifies drastically
- Recovers ℓ_2-Gromov-Wasserstein as particular case
- Application to unsupervised word translation yields SOTA-level performance at a fraction of the computational cost

Motivation

- Many problems in machine learning require correspondences between shapes/collections/point-clouds

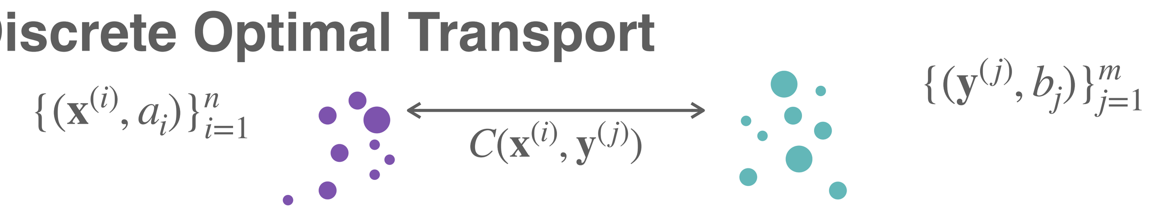


- Invariances common, especially on learnt representations
- Example: word embeddings across languages. Can we translate without supervision?



Background

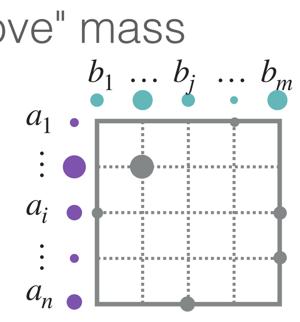
Discrete Optimal Transport



- Distance between distributions = cost to "move" mass

$$\min_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} \langle \Gamma, C(\mathbf{X},\mathbf{Y}) \rangle$$

$$= \{\Gamma \in \mathbb{R}^{n \times m} \ : \ \Gamma \mathbf{1} = \mathbf{a}, \Gamma^\top \mathbf{1} = \mathbf{b}\}$$



- Γ gives correspondences between points
- ... but requires spaces be globally aligned!

Orthogonal Procrustes Problem

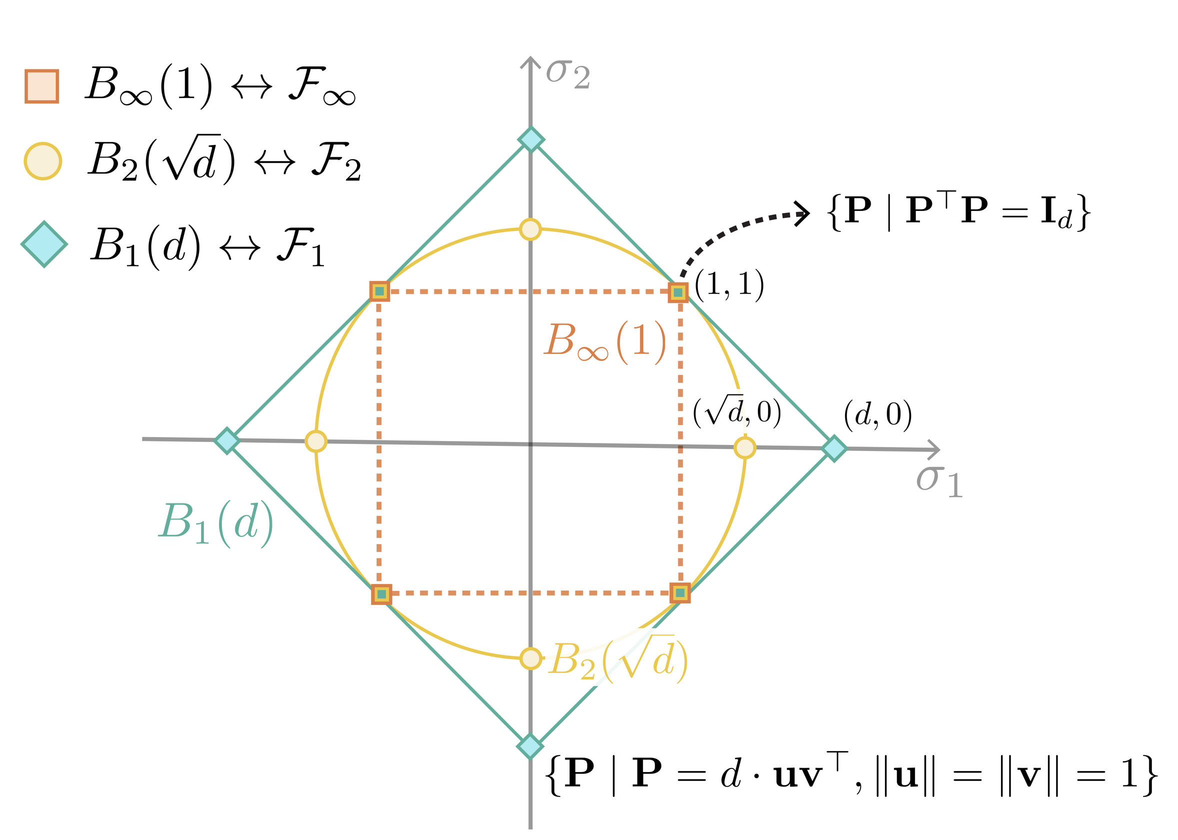
- Given known correspondences, find best rigid (orthogonal) mapping between them:

$$\min_{\mathbf{P} \in \mathfrak{O}(n)} \|\mathbf{X} - \mathbf{P}\mathbf{Y}\|_F^2$$

- Closed-from solution in terms of SVD of $\mathbf{X}\mathbf{Y}^\top$
- Easily generalized to other norms (Jaggi, 2013)

Schatten-Norm Invariances

- Invariance classes are functions of bounded Schatten norm:



- Modeling appeal. Clear interpretation:
  - $p = 1$: sparse spectra (projections) [Nuclear norm]
  - $p = 2$: radial spectra [Frobenius Norm]
  - $p = \infty$: uniform spectra (orthogonal) [Spectral norm]
- Algebraic + Computational convenience:
  - Unitary invariance
  - Submultiplicative
  - Easy characterization via duality

Transporting with Global Geometric Invariances

- Suppose $\exists f: \mathcal{Y} \to \mathcal{Y}$ s.t. $\forall \mathbf{x} \in \mathcal{X}, \exists \mathbf{y} \in \mathcal{Y}$ s.t. $\mathbf{x} \approx f(\mathbf{y})$
- We know invariance class $\mathcal{F}$, but not actual $f \in \mathcal{F}$
- Goal: find optimal coupling and global transform:

$$\min_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} \min_{f \in \mathcal{F}} \langle \Gamma, C(\mathbf{X}, f(\mathbf{Y})) \rangle$$

- We consider invariances classes of the form:

$$\mathcal{F}_p \triangleq \{\mathbf{P} \in \mathbb{R}^{d \times d} \mid \|\mathbf{P}\|_p \leq k_p\}$$

Schatten p-norm $\|\mathbf{P}\|_p = \|\sigma(\mathbf{P})\|_p$

Lemma [The objective simplifies]

For the euclidean ℓ_2 cost, if either of these holds:
(I) $\mathbf{P}$ is angle preserving
(II) $\mathbf{Y}$ is $\mathbf{b}$-whitened
Then, the problem is equivalent to:

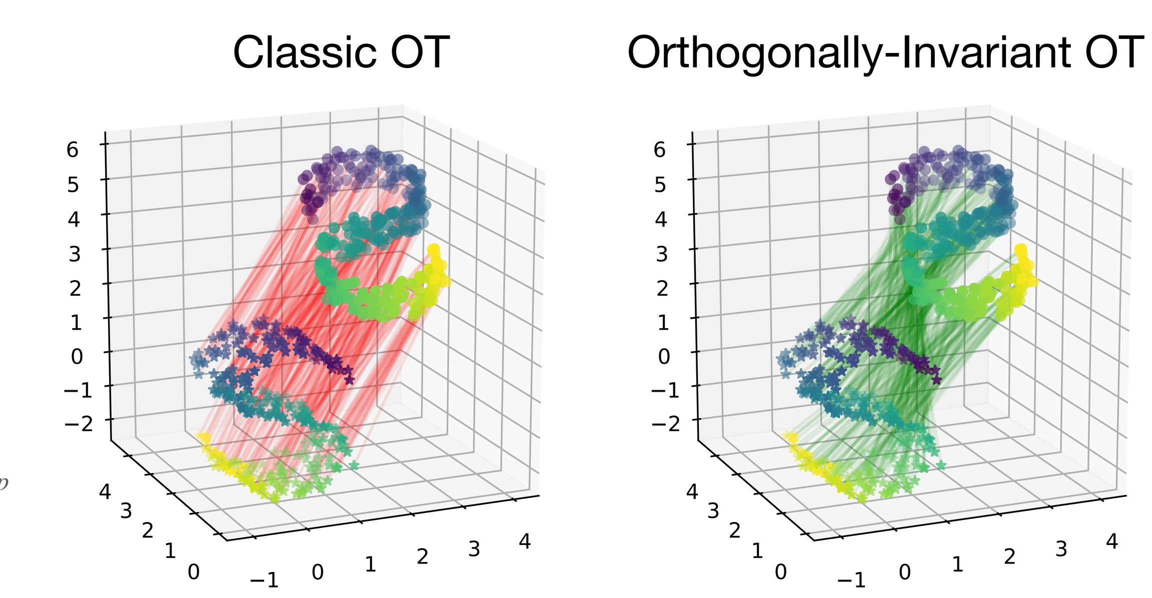
$$\underbrace{\max_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})}}_{\text{OT problem}} \underbrace{\max_{f \in \mathcal{F}} \langle \mathbf{X}\Gamma\mathbf{Y}^\top, \mathbf{P} \rangle}_{\text{Generalized Procrustes}}$$

$\boxed{\mathcal{F}\text{-Invariant Optimal Transport}}$

Lemma [Inner problem has closed-form sol.]

Let $\mathbf{U}\Sigma\mathbf{V}^\top$ be SVD decomposition of $\mathbf{X}\Gamma\mathbf{Y}^\top$, and let $\mathbf{s}$ be such that $\|\mathbf{s}\|_p \leq k$ and $\mathbf{s}^\top \sigma = k\|\sigma\|_q$. Then:

$$\arg\max_{f \in \mathcal{F}_p} \langle \mathbf{X}\Gamma\mathbf{Y}^\top, \mathbf{P} \rangle = \mathbf{U} \operatorname{diag}(\mathbf{s}) \mathbf{V}^\top$$



The case $p = \infty$

- Invariance to orthogonal transformations

$$\max_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} \max_{f \in \mathcal{F}_\infty} \langle \mathbf{X}\Gamma\mathbf{Y}^\top, \mathbf{P} \rangle = \max_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} \|\mathbf{X}\Gamma\mathbf{Y}^\top\|_*$$

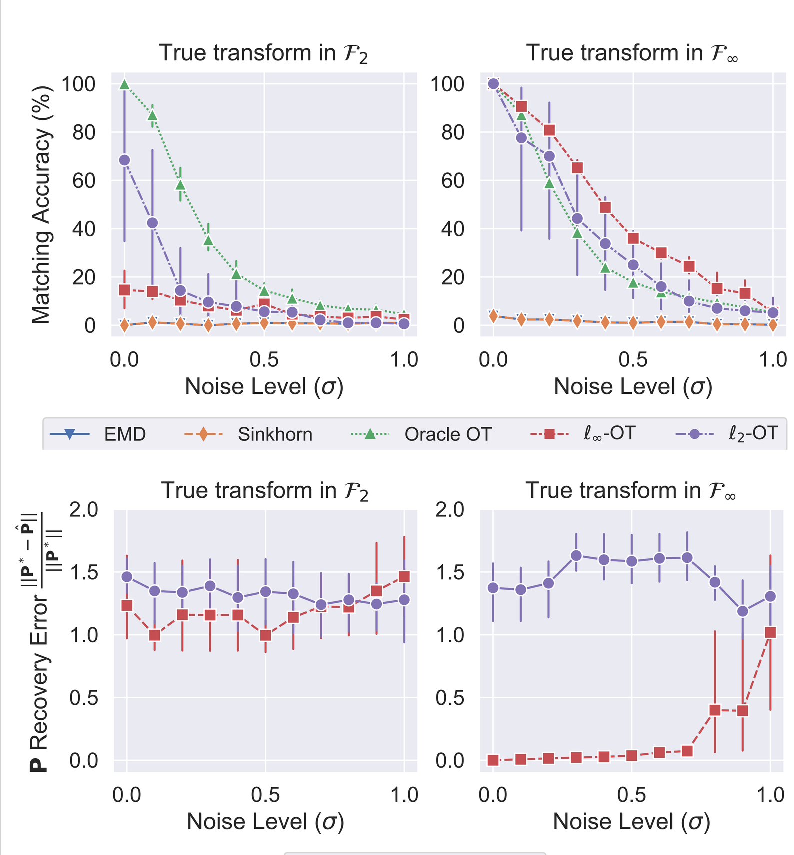
Optimization

- The problem is not jointly concave. Can solve by alternating optimization, but sensitive to initialization.
- Entropy regularization (Cuturi, 2013) allows efficient solution of outer problem + controls non-concavity
- Annealing regularization makes algorithm robust to initialization!

Experiments

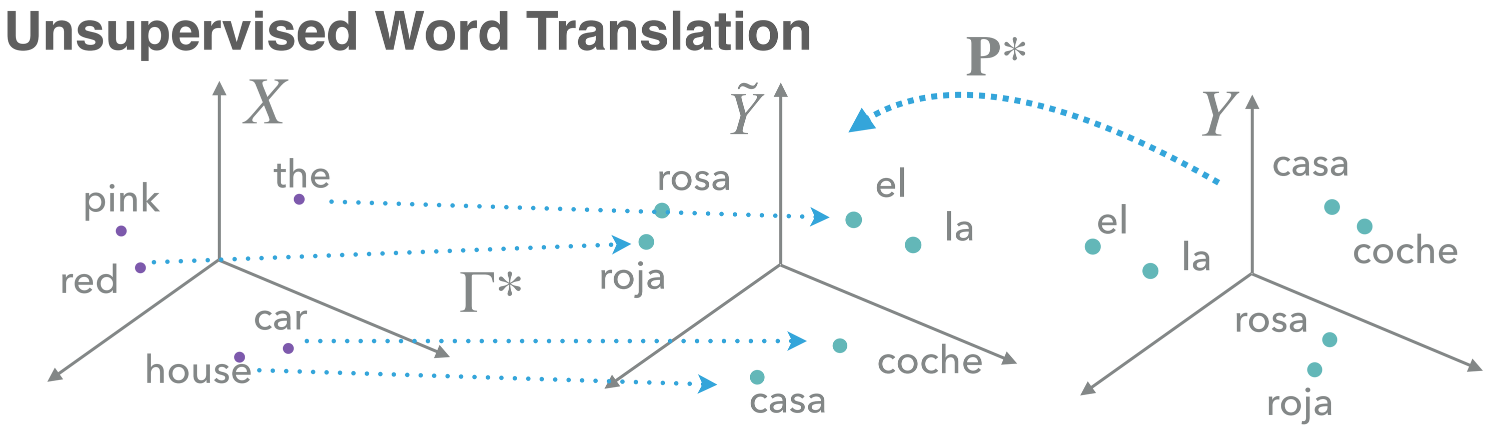
Synthetic

- Known latent transformation
- S-shape point cloud in $\mathbb{R}^3$
- Random transform from $\mathcal{F}_2$ or $\mathcal{F}_\infty$
- Two metrics of interest:
  - Recovery of F
  - Accuracy of matching points

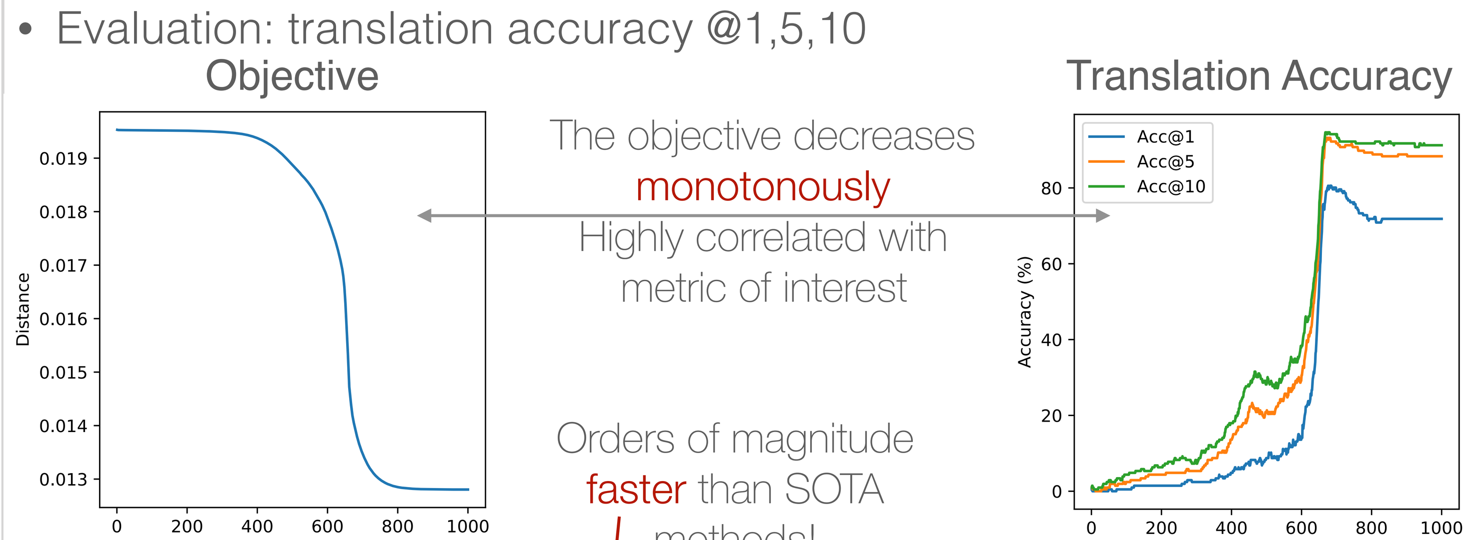


- TLDR: Better recovery if optimizing over correct invariance class.

Unsupervised Word Translation



- FastText word embeddings, 6 language pairs
- Invariance: orthogonal transforms
- Evaluation: translation accuracy @1,5,10



	Supervision	Time	En-Es →	En-Es ←	En-Fr →	En-Fr ←	En-De →	En-De ←	En-It →	En-It ←	En-Ru →	En-Ru ←
Procrustes	5K words	3	77.6	77.2	74.9	75.9	68.4	67.7	73.9	73.8	47.2	58.2
Procrustes + Csls	5K words	3	81.2	82.3	81.2	82.2	73.6	71.9	76.3	**75.5**	51.7	63.7
Adv + Csls	None	643	75.7	79.7	77.8	71.2	70.1	66.4	72.4	71.2	37.1	48.1
Adv + Csls + Refine	None	957	81.7	83.3	82.3	82.1	74.0	72.2	77.4	76.1	44.0	59.1
Wasserstein + Csls	None	–	82.8	84.1	82.6	82.9	75.4	73.3	–	–	43.7	59.1
Gromov-Wasserstein	None	37	81.7	80.4	81.3	78.9	71.9	72.8	78.9	75.2	45.1	43.7
Self-Learn + Csls	None	476	82.3	84.7	82.3	83.6	75.1	74.3	79.2	79.8	48.9	65.9
ℓ_∞-InvarOt + Csls	None	70	81.3	81.8	82.9	81.6	73.8	71.1	77.7	77.7	41.7	55.4

Connection to Gromov-Wasserstein

- Take $p = 2$ (Frobenius norm invariance)

$$\mathcal{F}_2 = \{\mathbf{P} \in \mathbb{R}^{d \times d} \mid \|\mathbf{P}\|_F \leq \sqrt{d}\}$$

- The problem becomes:

$$\max_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} \max_{f \in \mathcal{F}_2} \langle \mathbf{X}\Gamma\mathbf{Y}^\top, \mathbf{P} \rangle = \sqrt{d} \max_{\Gamma \in \Pi(\mathbf{a},\mathbf{b})} \|\mathbf{X}\Gamma\mathbf{Y}^\top\|_F$$

Lemma

This is equivalent to computing the Gromov-Wasserstein distance (Memoli, 2011):

$$\min_{\Pi(\mathbf{a},\mathbf{b})} \sum_{i,j,k,l} L(\mathbf{C}^x_{ik}, \mathbf{C}^y_{jl}) \Gamma_{ij} \Gamma_{kl}$$

where $\mathbf{C}^x, \mathbf{C}^y$ are similarity matrices in the ℓ_2 metric, and L is the ℓ_2 loss.

Future Work

- Relaxing assumptions - requires inner optimization too, solvable with Frank-Wolfe
- Alternative optimization via level-set methods
- Dual OT view of the problem
- Applications with non-orthogonal invariances

Key References

(1) Conneau et al. "Word Translation Without Parallel Data", ICLR 2018.
(2) Cuturi, M. " Sinkhorn distances: Lightspeed computation of optimal transport", NIPS 2014
(3) Jaggi, M. "Revisiting Frank-Wolfe: Projection-free sparse convex optimization", ICML 2013.
(4) Memoli, F. "Gromov–Wasserstein distances and the metric approach to object matching", FCM 2011.