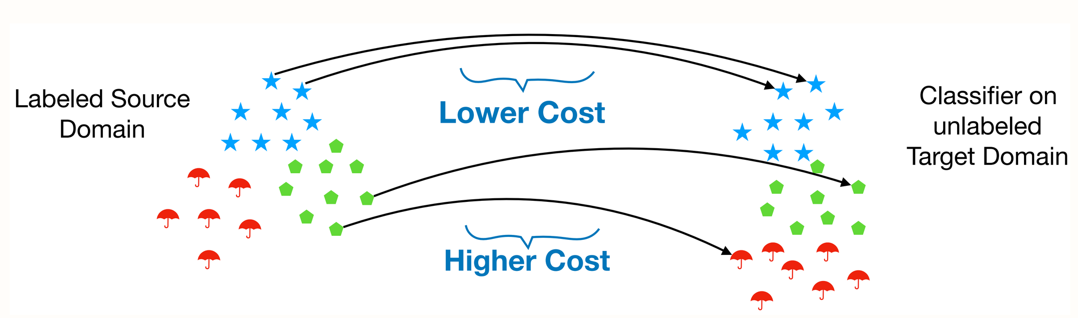
Structured Optimal Transport

David Alvarez-Melis, Tommi S. Jaakkola, Stefanie Jegelka | Massachusetts Institute of Technology

Summary

- A general framework for injecting structure into OT
- Submodularity offers flexibility + tractability (via convexity)
- Fast algorithms via saddle-point and convex optimization
- Applications to domain adaptation, sentence similarity

Motivation



- Can we inject structure into the cost definition of OT?
- Should remain tractable (~convex)

Background

Discrete Optimal Transport

- Discrete distributions: $\mu = \sum_{i=1}^{n} p_i^s \delta_{x_i^s}$, $\nu = \sum_{i=1}^{m} p_i^t \delta_{x_i^t}$
- Ground cost matrix $C_{ij} = C(x_i^s, x_j^t)$.
- **Transport polytope**: $\mathcal{M}_{\mu,\nu} = \{\gamma \in \mathbb{R}_+^{n\times m} \mid \gamma \mathbf{1} = \mu,\ \gamma^T \mathbf{1} = \nu\}$

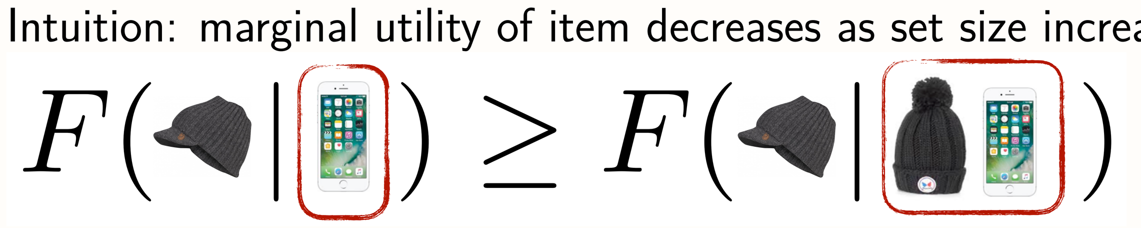
  The Problem: $\min_{\gamma \in \mathcal{M}_{\mu,\nu}} \sum_{i,j} \gamma_{ij} C_{ij}.$
- Objective is separable in γ_{ij}: no interaction between assignments!!

Submodularity

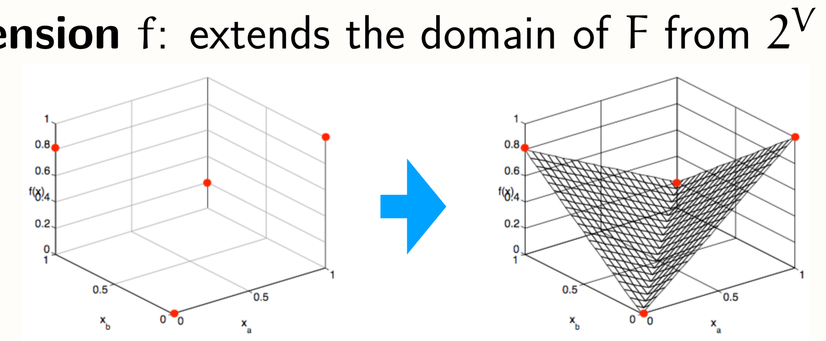
- Set function $F : 2^V \to \mathbb{R}$ is **submodular** if:

$$F(S \cup \{v\}) - F(S) \geq F(T \cup \{v\}) - F(T) \quad \forall S \subseteq T, v \notin T$$

- Analogues to convexity/concavity
- Intuition: marginal utility of item decreases as set size increases



- **Lovász Extension** f: extends the domain of F from 2^V to $\mathbb{R}_+^n$



- f is convex iff F is submodular
- For F submodular, $f(w) = \max_{x \in \mathcal{B}_F} w^T x$
- **Base polytope** $\mathcal{B}_F$ is "nice", leads to tractability ☺

Approach

OT with submodular costs

- Discrete (matching) view of OT (~Monge formulation)
- Matching with submodular costs:

$$F(M) = \sum_{\ell} g_l \Big(\sum_{(i,j) \in M \cap G_\ell} c_{ij} \Big), \qquad g \text{ concave}$$

- E.g., $g_l(x) = \min\{x, \alpha\} + \sqrt{[x-\alpha]_+}$
- Want continuous, fractional assignments
- Relax objective to Lovasz extension!

$$\min_{\gamma \in \mathcal{M}} f(\gamma) \equiv \min_{\gamma \in \mathcal{M}} \max_{\kappa \in \mathcal{B}_F} \langle \gamma, \kappa \rangle$$

Optimization

$\min_{\gamma \in \mathcal{M}} f(\gamma)$

- Non-smooth, convex
- Mirror Descent: $O(\frac{1}{\sqrt{t}})$

$\min_{\gamma \in \mathcal{M}} \max_{\kappa \in \mathcal{B}_F} \langle \gamma, \kappa \rangle.$

- **Smooth** convex-concave
- Saddle-Point Mirror-Prox: $O(\frac{1}{t})$

Subroutines

Subgradients of f

- Subdifferential of f: $\partial f(\gamma) = \operatorname{argmax}_{\kappa \in \mathcal{B}_F} \langle \kappa, \gamma \rangle$.
- Linear optimization over base polytope
- Solved by Edmond's greedy algorithm (~sorting) in $O(N \log N)$

Projections on $\mathcal{M}$

- Entropic mirror map $\Phi_{\mathcal{M}}(\gamma) := \sum_{i,j} \gamma_{ij} \ln(\gamma_{ij})$ yields:

$$\hat{\gamma} = \operatorname*{argmin}_{\gamma \in \mathcal{M}} \mathrm{KL}(\gamma \parallel w).$$

- Solved with Sinkhorn-Knopp [1].

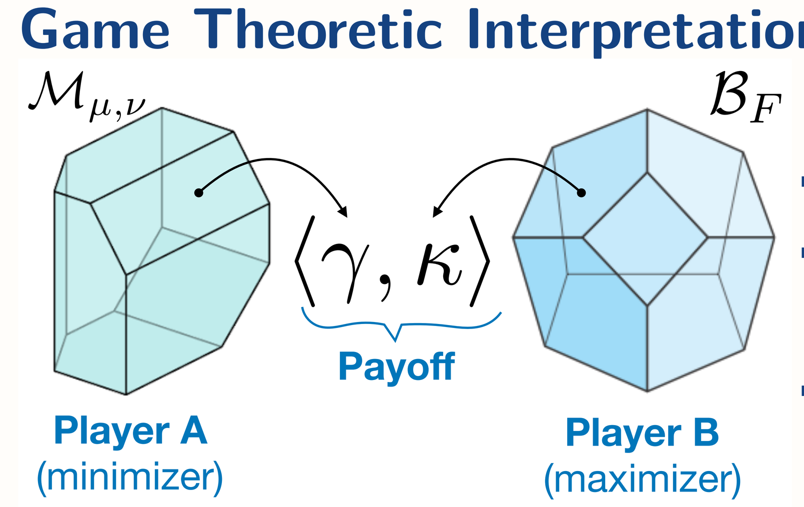
Projections on $\mathcal{B}_F$

- Euclidean mirror map $\Phi_{\mathcal{B}_F}(\kappa) = \frac{1}{2}\|\kappa\|^2$ yields:

$$\hat{\kappa} = \operatorname*{argmin}_{\kappa \in \mathcal{B}_F} \|\kappa - w\|_2^2$$

- Solved e.g. via the Fujishige-Wolfe minimum norm point algo
- For our *decomposable* functions, can do in $O(|E| \log |E|)$
- If F_i have disjoint supports, compute projections in parallel
- If not, randomized coordinate descent [2]

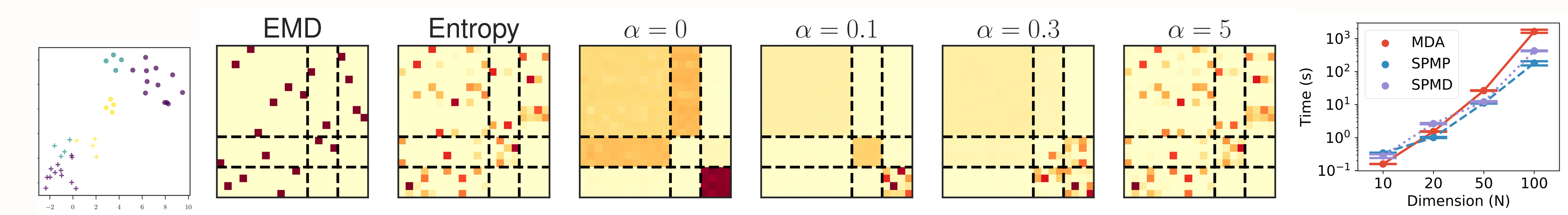
Game Theoretic Interpretation



- ∃ Nash equilibrium
- "Power of B" ≈ submodularity strength
- Classic OT ⇔ $|\mathcal{B}_F| = 1$

Experiments

Clustered Point Cloud Matching

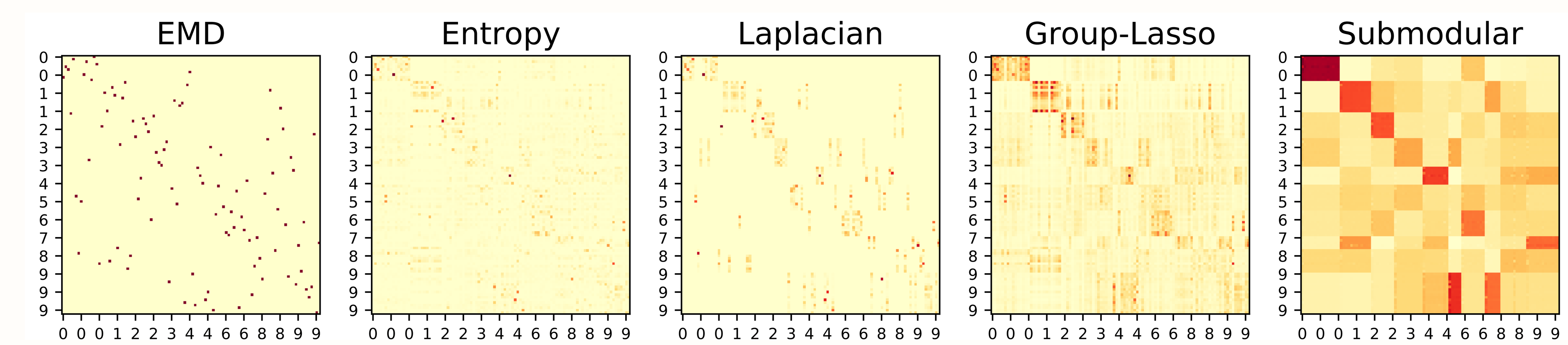


Small α: aggressive cluster enforcement. Large α: recovers entropy-regularized solution

Domain Adaptation

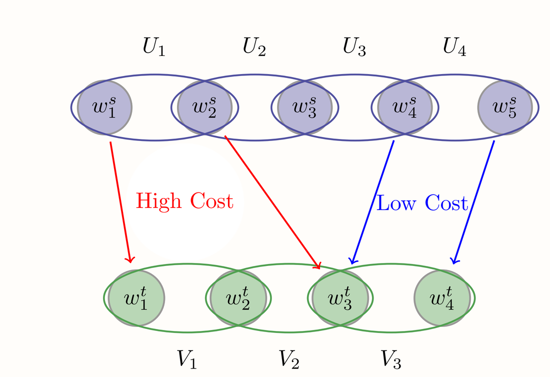
- Objective: encourage points of the same class to be mapped together
- [3] use penalty-based methods
- Task: USPS ↔ MNIST digit adaptation
- $N_s = N_t = 100$ examples.

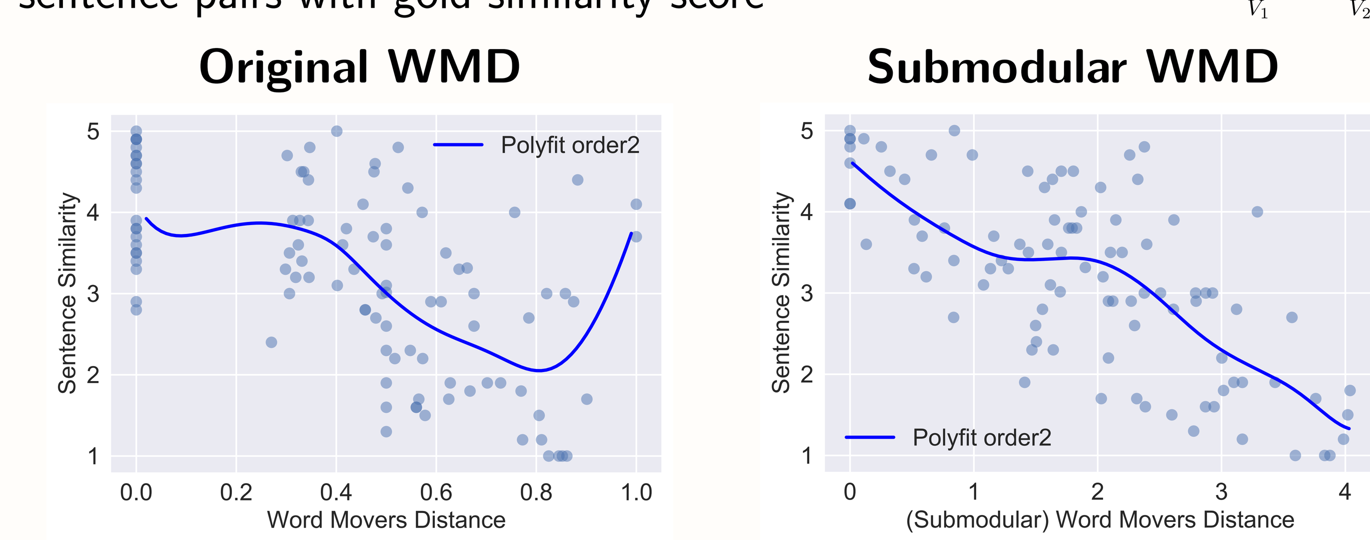
Method	MNIST→USPS	USPS→MNIST
No adaptation	41.20	33.10
EMD	37.72	33.68
Entropy	55.70	43.64
Laplace	54.37	37.73
Group-Lasso	57.12	49.49
SOT	**62.97**	**58.34**



Sentence Similarity

- Word mover's distance [4] measures sentence similarity
- Ground metric: distances between word embeddings
- WMD ignores positions of words in sentence
- SOT allows for a syntax-aware version of the WMD
- SICK dataset: sentence pairs with gold similarity score





MSE 0.67 (Spearman's $\rho = .71$) MSE=0.59 (Spearman's $\rho = .75$)

Future Work

- Other structures (trees, hierarchies)
- Beyond submodularity
- Speed-up by stochastic optimization
- Use in Generative Adversarial Nets

Key References

[1] M. Cuturi. "Sinkhorn distances: Lightspeed computation of optimal transport". In: *NIPS*. 2013.

[2] A. Ene and H. L. Nguyen. "Random Coordinate Descent Methods for Minimizing Decomposable Submodular Functions". In: *ICML*. 2015.

[3] N. Courty et al. "Optimal Transport for Domain Adaptation". In: *TPAMI* (2017).

[4] M. J. Kusner et al. "From Word Embeddings To Document Distances". In: *ICML* 37 (2015), pp. 957–966.

Contact: {davidam|tommi|stefje}@csail.mit.edu