Structured Optimal Transport: Supplementary Material

David Alvarez-Melis
MIT CSAIL

Tommi S. Jaakkola
MIT CSAIL

Stefanie Jegelka
MIT CSAIL

A The structured optimal transport is a semi-metric

We restate Lemma 3.1 and prove it.

Lemma A.1. *Suppose the ground cost $c(\cdot,\cdot)$ is a metric and that F is a submodular non-decreasing function such that $F(\emptyset) = 0$ and $F(\{(i,j)\}) > 0$ iff $c(x_i, y_j) > 0$. Then $d_F(\mu,\nu) = \min_{\gamma\in\mathcal{M}} f(\gamma)$ is a semi-metric.*

Proof. Let $\mathbf{C} \in \mathbb{R}^{n\times m}$ be the cost matrix associated with c, i.e. $\mathbf{C}_{ij} = c(x_i, y_j)$ for $i \in \{1,\dots,n\}$ and $j \in \{1,\dots,m\}$. In addition, define $\mathbf{p}$ and $\mathbf{q}$ to be the vectors of probability weights of μ and ν, respectively, i.e. $\mu = \sum_i^n p_i\delta_{x_i}$ and $\nu = \sum_j^m q_j\delta_{y_j}$.

Since $c(\cdot,\cdot)$ is a metric, every $\mathbf{C}_{ij}$ is non-negative. Furthermore, since we assume support points are not duplicated, $\mathbf{C}$ has at most n zero entries, and the rest are strictly positive. This, combined with the fact that F is non-decreasing, implies $F(S) \geq 0$ for every $S \subseteq V$, and therefore its Lovász extension must also be non-negative. In particular,

$$d_F(\mu,\nu) = \min_{\gamma\in\mathcal{M}} f(\gamma) \geq 0 \qquad \forall \mu,\nu \tag{18}$$

Now, suppose $\mu = \nu$, and without loss of generality, assume the support points are indexed such that $x_i = y_i$ for every i. In addition, we must have $\mathbf{p} = \mathbf{q}$, so $\gamma = \text{diag}(\mathbf{p}) \in \mathcal{M}$. On the other hand, since c is a metric $\mathbf{C}_{ii} = 0$ for every i, which in turn implies that for any $\kappa \in \mathcal{B}_F$ and every i, $\kappa_{ii} \leq F(\{i,i\}) = 0$. By (18) and the minimax equilibrium properties, we have

$$0 \leq d_F(\mu,\nu) = \langle\gamma^*,\kappa^*\rangle \leq \langle\gamma,\kappa^*\rangle \qquad \forall\gamma\in\mathcal{M}$$

In particular, for $\gamma = \text{diag}(\mathbf{p})$, we get

$$0 \leq d_F(\mu,\nu) \leq \sum_i p_i\kappa^*_{ii} \leq 0$$

So we conclude that $d_F(\mu,\nu) = 0$. Conversely, let $d_F(\mu,\nu) = 0$, and suppose, for the sake of contradiction, that $\mu \neq \nu$. Then, at least one of the following is true:

(i) $\mathbf{p} \neq \mathbf{q}$

(ii) the support points are different, i.e. there is no reordering of indices such that $x_i = y_i$ for every i.

If (i) is true, $\mathcal{M}$ cannot be a permutation matrix, so in particular γ^* has at least $n+1$ positive entries. We can thus find a $\kappa \in \mathcal{B}_F$ which has positive weights in all those entries. In that case, we have $\langle\gamma^*,\hat{\kappa}\rangle > 0$, a contradiction. Now, if on the other hand (ii) is true, then $\mathbf{C}$ has strictly less than n zero entries. This, by our assumptions on F, means that there exist $\kappa \in \mathcal{B}_F$ with less than n non negative entries. Any such matrix will have $\langle\gamma^*,\kappa\rangle > 0$, a contradiction.

Finally, the symmetry of $d_F(\mu,\nu)$ is trivial. □

B Topological constraints in Structured Optimal Transport

Besides the settings presented in this work where structure arises from group labels, the framework proposed here allows us to explicitly encourage certain topological aspects of the distributions to be preserved. One such possible constraint for discrete distributions that lie on a low-dimensional manifold is to encourage neighboring points to be matched together. Such type of constraints can substantially alter the resulting transport plans, as shown in Figure 5 for a simple two-moons dataset. Here, the SOT solution favors neighborhood preservation over element-wise cost, resulting in a block-structured optimal coupling.

C The Sinkhorn-Knopp Matrix Scaling Algorithm

Cuturi (2013) proposes to solve the entropy-regularized optimal transport problem

$$\underset{\gamma\in\mathcal{M}}{\text{argmin}}\langle\gamma,C\rangle - \frac{1}{\lambda}H(\gamma) \tag{19}$$

with the Sinkhorn-Knopp matrix scaling algorithm. Lemma 2 in (Cuturi, 2013), based on Sinkhorn's Theorem (Sinkhorn, 1967), shows that there exists a unique solution to this problem, and that it has the form

$$\gamma^*_\lambda = \text{diag}(u)\mathbf{K}\text{diag}(v)$$

where $\mathbf{K}$ is the entry-wise exponential of $-\frac{1}{\lambda}C$ and $u,v \in \mathbb{R}^d_+$. Furthermore, u and v can be efficiently

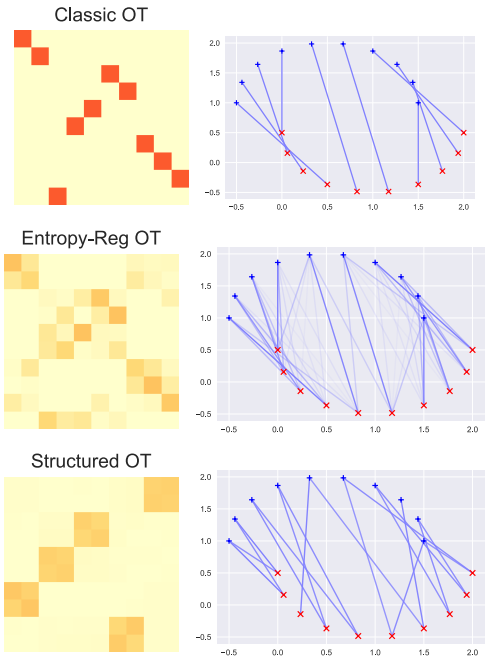

Figure 5: Optimal transport plans and matchings for the two moons example.

obtained by means of Sinkhorn's fixed-point iteration, which involves updates of the form:

$$u^{(n+1)} = \mu \,./(Kv^{(n)})$$
$$v^{(n+1)} = \nu \,./(K^T u^{(n)})$$

where, again, the division is entry-wise. The iterates $u^{(n)}$ and $v^{(n)}$ converge linearly to the true u and v.

D Fast projections into submodular function base polytopes

The problem of computing the point of minimal norm on the base polytope of a submodular function is intimately related to that of minimizing the function itself. The solutions to these two problems are related through the parametric minimization problem

$$S^*_\lambda = \operatorname{argmin} F(S) - \lambda|S|$$

Let $\mathbf{y}^*$ be the min-norm point in $\mathbf{B}_F$. We can recover the solution to the original submodular function minimization (SFM) problem, $S^* := S^*_{\lambda=0}$ from y^* as $S^* = \{i \mid y^*_i \leq 0\}$. Conversely, we can recover $\mathbf{y}^*$ from the solutions of the parametric problem as

$$\mathbf{y}^*_j = \max\{\lambda \mid j \in S^*_\lambda\}$$

Given a method for minimizing the function $F^\lambda := F(S) - \lambda|S|$, one can obtain the min-norm-point by repeated calls to this oracle and a divide-and-conquer strategy as the one Jegelka, Bach, and Sra (2013) use, which runs in $O(n \log n)$ time.

Now, in our case, we are dealing with cluster functions of the form $F_i(S) = g(\sum_{i\in S} w_i)$, and in addition, we are interested in computing projections, rather than the min-norm-point, i.e., we are interested in $\tilde{\kappa} = \operatorname{argmin}_{\kappa\in\mathcal{B}_F} \|\kappa - m\|_2^2$ for some $m \in \mathbb{R}^{n\times m}$. Equivalently, we want to minimize $F_w(S) := F(S) - M(S)$, where M is the modular function implied by the vector m. Thus, the parametric submodular function minimization (SFM) problem we are dealing with is

$$\begin{aligned}
F^\lambda_w &= g(\sum_{i\in S} w_i) + \sum_{i\in S} m_i - \lambda|S| \\
&= g(\sum_{i\in S} w_i) + \sum_{i\in S} (m_i - \lambda) \\
&= \min_{\alpha\in I} c_\alpha + (\alpha \sum_{i\in S} w_i) + \sum_{i\in S} (m_i - \lambda) \\
&= \min_{u\in[0,\sum_{i\in V} w_i]} g(u) + \nabla g(u)\big(\sum_{i\in S} w_i - u\big) + \sum_{i\in S} (m_i - \lambda)
\end{aligned}$$

where we used the fact that any concave function can be written as the pointwise supremum of (potentially infinite) linear functions, parametrized by α, and an interval I where the valid gradients lie. Since the minimization is jointly over S and α, we can rewrite the problem as

$$\min_\alpha \min_S c_\alpha + \alpha \sum_{i\in S} w_i + \sum_{i\in S} (m_i - \lambda) \tag{20}$$

As the slope $\alpha = \nabla g(u)$ shrinks, the constant $c_\alpha = g(u) - u\nabla g(u)$ grows. We make the following observations:

1. Equation (20) suggests the following strategy: (1) for each α, find the minimizing set S^α. (2) Evaluate the function above for each S^α, and pick the one minimizing $F(S)$.

2. For a fixed α, the optimal S^α is easy to find:

   $$S^\alpha - \{i \mid \alpha w_i + m_i + \lambda \leq 0\} = \{i \mid \alpha \leq -(m_i + \lambda)/w_i$$

3. Observation 2 shows that the optimal sets as α shrinks are nested: once an item enters the optimal set, it never leaves.

These observations suggest a simple sorting-based algorithm for finding the minimizer of $F(S)$, shown here as Algorithm 3. It runs in time $O(n \log n + nT)$, where T is the evaluation time of F and n is the size of the ground set of F. We emphasize that this algorithm is only valid for the concave-of-sum functions as defined in Section 3.1.

Algorithm 3 Fast SFM for Concave-of-Sum

Input: Initial point $z^0 = (\gamma_0, \kappa_0)$ and step size η_0
for $i = 1, \dots, n$ **do**
  $r_i \leftarrow -(m_i + \lambda)/w_i$
end for
$\hat{V} \leftarrow Sort(V)$ {By value of r_i}
for $k = 1, \dots, n$ **do**
  $S_k \leftarrow \{1, \dots, V(k)\}$
end for
$S^* = \operatorname{argmin}_{S_i} F(S_i)$
return S^*

E Edmond's sorting algorithm

Let f be the Lovász extension of a submodular function $F : 2^V \to \mathbb{R}$. Then f can be evaluated at $w \in \mathbb{R}^n$ as follows. Let σ be a reordering of the elements of V such that $w_{\sigma_1} \geq w_{\sigma_2} \geq \cdots \geq w_{\sigma_n}$, and define $S_i = \{\sigma_1, \dots, \sigma_i\}$. Then

$$f(w) = \sum_{i=1}^{n} w_{\sigma_i} \big[F(S_i) - F(S_{i-1})\big]$$

The computational cost in this procedure is dominated by the sorting. Now, recalling that equivalence $f(x) = \max_{y \in \mathcal{B}_F} \langle y, x \rangle$, we note that this same procedure yields the maximizing y, setting $y_{\sigma_i} := F(S_i) - F(S_{i-1})$. It is trivial to verify that indeed $y \in \mathcal{B}_F$.

F Derivation of Mirror Descent Steps

We derive here the steps for SP-MD. The derivation for MDA (Algorithm 1) and SP-MP (Algorithm 2) is analogous.

Let $\mathcal{Z} = \mathcal{M} \times \mathcal{B}_F$, and denote by $z \in \mathcal{Z}$ a pair $z = (\gamma, \kappa)$. Suppose $\Phi_{cM}, \Phi_{\mathcal{B}}$ are mirror maps on $\mathcal{M}$ and $\mathcal{B}_F$, respectively. We define $\Phi_{\mathcal{Z}}(z = (\gamma, \kappa)) := \Phi_{\mathcal{M}}(\gamma) + \Phi_{\mathcal{B}}(\kappa)$. The SP-MD algorithm computes at every step:

a) $w_{t+1} \in D$ such that $\nabla\Phi(w_{t+1}) = \nabla\Phi(z_t) - \eta g_t$

b) $z_{t+1} \in \operatorname{argmin}_{z \in \mathcal{Z}} D_\Phi(z, w_{t+1})$

Note that $\Phi = (\Phi_{\mathcal{M}}, \Phi_{\mathcal{B}})$, so (a) amounts to finding $w_{t+1} = (w^\gamma_{t+1}, w^\kappa_{t+1})$ such that:

$$\nabla\Phi_{\mathcal{M}}(w^\gamma_{t+1}) = \nabla\Phi_{\mathcal{M}}(\gamma_{t+1}) - \eta\kappa_t \tag{21}$$
$$\nabla\Phi_{\mathcal{B}}(w^\kappa_{t+1}) = \nabla\Phi_{\mathcal{B}}(\kappa_{t+1}) + \eta\gamma_t \tag{22}$$

At this point, the updates take different forms depending on the mirror maps. For our choice of $\Phi_{\mathcal{M}}(\gamma) = H(\gamma)$, we have $\nabla\Phi_{\mathcal{M}}(\gamma) = \mathbf{1} + \log\gamma$ (where the logarithm is to be understood element-wise), so (21) becomes:

$$\log w^\gamma_{t+1} = \log\gamma_t - \eta\kappa_t \tag{23}$$

Algorithm 4 Saddle Point Mirror Descent for Structured Optimal Transport

Input: Initial point $z^0 = (\gamma_0, \kappa_0)$ and step size η_0
while $\epsilon_{SP} < tol$ **do**
  $\gamma_{t+1} \leftarrow$ SINKHORN$(\gamma_t \circ \exp\{-\eta_t\kappa_t\})$
  $\kappa_{t+1} \leftarrow$ BASEPOLYPROJECT$(\kappa_t + \eta_t\gamma_t)$
  $z^{t+1} \leftarrow [\sum_{s=1}^{t+1}\eta_s]^{-1}\sum_{s=1}^{t+1}\eta_s(\gamma_s, \kappa_s)$
  $\epsilon_{SP} \leftarrow$ SADDLEGAP(z^t)
  $t \leftarrow t + 1$
end while

Hence,

$$w^\gamma_{t+1} = \gamma_t \cdot e^{\eta\kappa_t},$$

where the product and exponential are, again, element-wise. On the other hand, for the mirror map $\Phi_{\mathcal{B}}(\kappa) = \frac{1}{2}\|\kappa\|_2^2$, (22) becomes

$$w^\kappa_{t+1} = \kappa_t + \eta\gamma_t \tag{24}$$

The second step in SPMD (step (b) above) requires projecting w_{t+1} and thus $(w^\gamma_{t+1}, w^\kappa_{t+1})$ into $(\mathcal{M}, \mathcal{B}_F)$ according to the Bregman divergences associated with the mirror maps $\Phi_{\mathcal{M}}(\gamma), \Phi_{\mathcal{B}}(\kappa)$. For the entropy map, this becomes an KL-divergence projection, so we have

$$\gamma_{t+1} \in \operatorname*{argmin}_{\gamma} \mathrm{KL}(\gamma \parallel \gamma^t \cdot e^{\eta\kappa_t}) \tag{25}$$

On the other hand, the divergence associated with the ℓ_2 norm map is again an ℓ_2 distance, so

$$\kappa_{t+1} \in \operatorname*{argmin}_{\kappa} \|\kappa - \kappa_t + \eta\gamma_t\|_2^2 \tag{26}$$

The full SP-MD Algorithm is shown as Algorithm 4.

G Shortcomings of the Word Mover's Distance

There are obvious limitations the WMD's purely semantic bag-of-words approach to sentence similarity, arising from ignoring the relations among words in a sentence. For example, consider the following sentences:

a) The hotel does not appear in this book
b) I will book this hotel
c) I will reserve this hotel

The WMD between (a) and (b) will likely be less that than between (b) and (c), even though the latter two are paraphrases of each other. Although (a) and (b) have strong single-word semantic overlap, the order in which the words occur in these two sentences entails different meanings. As contrived as this example might be, it is a good reminder that syntax and word-meaning go hand-in-hand for assessing semantic similarity at the sentence level.

H Digit transportation

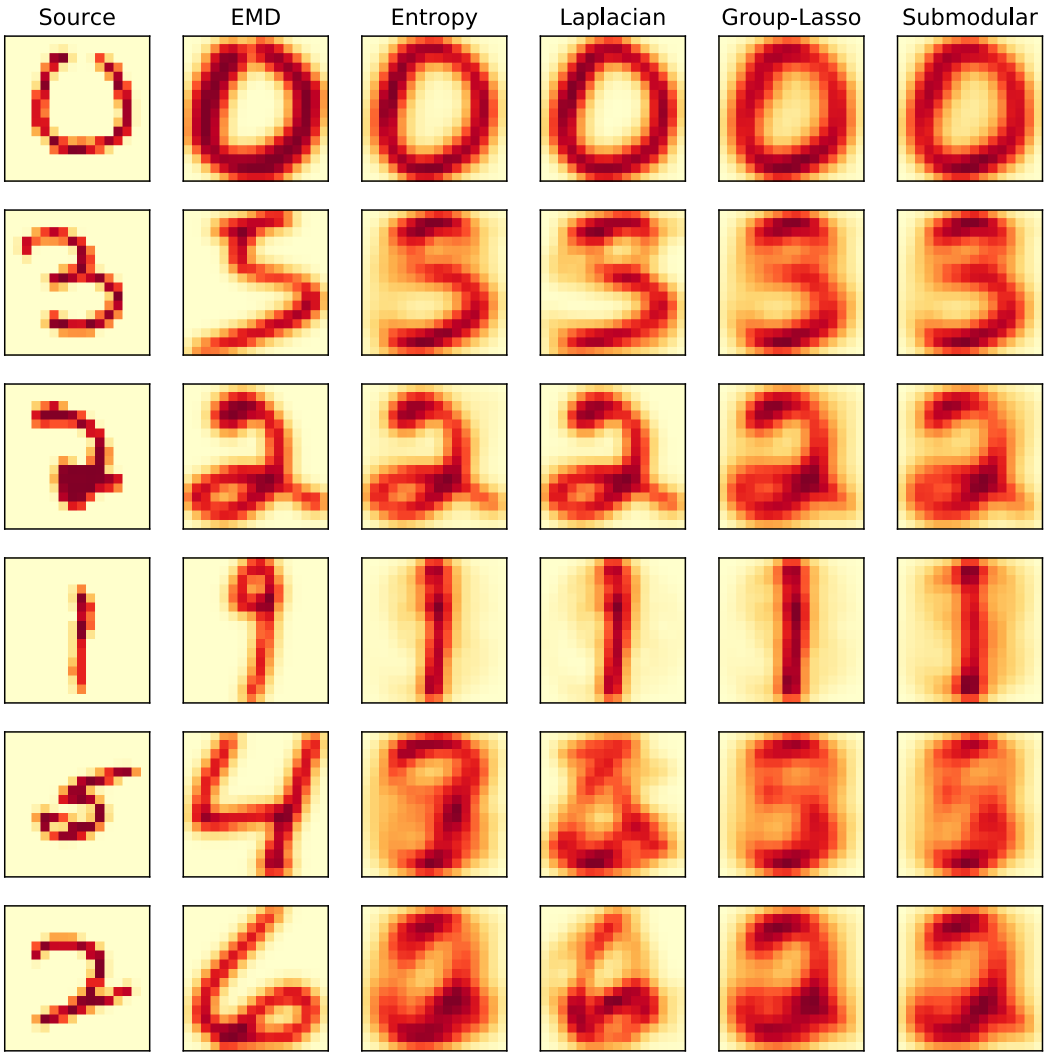

Figure 6: Examples from the MNIST→USPS domain adaptation task. The first column is the source image from MNIST, and the remaining columns are the result of transporting the source image into the target domain with the barycentric mapping defined by the various optimal transport plans.