Structured Optimal Transport

David Alvarez-Melis
MIT CSAIL

Tommi S. Jaakkola
MIT CSAIL

Stefanie Jegelka
MIT CSAIL

Abstract

Optimal Transport has recently gained interest in machine learning for applications ranging from domain adaptation to sentence similarities or deep learning. Yet, its ability to capture frequently occurring structure beyond the "ground metric" is limited. In this work, we develop a nonlinear generalization of (discrete) optimal transport that is able to reflect much additional structure. We demonstrate how to leverage the geometry of this new model for fast algorithms, and explore connections and properties. Illustrative experiments highlight the benefit of the induced structured couplings for tasks in domain adaptation and natural language processing.

1 Introduction

Optimal transport provides a natural, elegant framework for comparing probability distributions while respecting the underlying geometry (Villani, 2008). Due to its strong theoretical foundations and many desirable properties, both the continuous and discrete versions of the transportation problem have received considerable attention in various fields within and beyond mathematics, including statistics (Mallows, 1972), differential equations (Jordan, Kinderlehrer, and Otto, 1998), optics (Glimm and Oliker, 2003) and economics (Galichon, 2016). Within machine learning and related fields, optimal transport distances (in particular the Wasserstein metric) have found successful application to shape analysis (Gangbo and McCann, 2000), image registration and interpolation (Solomon et al., 2015), domain adaptation (Courty et al., 2017), adversarial neural networks (Arjovsky, Chintala, and Bottou, 2017), and multi-label prediction (Frogner et al., 2015).

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

The discrete version of the problem has also had impact in settings where relaxed notions of matchings are sought, such as pairing control and treatment units in observational studies (Rosenbaum and Rubin, 1985). The range of applications has been growing with the development of fast algorithms (Cuturi, 2013; Genevay et al., 2016).

An important appeal of optimal transport distances is that they reflect the metric of the underlying space in the transport cost. Yet, in a number of settings, there is further important structure that remains uncaptured. This structure can be *intrinsic* if the distributions correspond to structured objects (e.g., images with segments, or sequences) or *extrinsic* if there is side information that induces structure (e.g., groupings). A concrete example arises when applying optimal transport to domain adaptation, where a subset of the source points to be matched have known class labels. In this case, we may desire source points with the same label to be matched coherently to the same compact region of the target space, preserving compact classes, and not be split into disjoint, distant locations (Courty et al., 2017). When pairing control and treatment units in observational studies of treatment effects, it is beneficial to compare treated and control subjects from the same "natural block" (e.g., family, hospital) so as to minimize the difference between unmeasured covariates (Pimentel et al., 2015). In all these examples, the additional structure essentially seeks correlations in the mappings of "similar" source points. Such dependencies, however, cannot be induced by standard formulations of optimal transport whose cost is separable in the mapping variables;[1] they require nonlinear interactions.

In this work, we develop a framework to incorporate such structural information directly into the optimal transport problem. This novel formulation opens avenues to a much richer class of (nonlinear) cost functions, allowing us to encode known or desired interac-

[1] The original optimal transport formulation with cost $\sum_{ij} c_{ij}\gamma_{ij}$ is linear in the mappings γ_{ij}, γ_{kl} of separate source locations i, k; the mappings are counted independently.

tions of mappings, such as grouping constraints, correlations, and explicitly modeling topological information that is present, for instance, in sequences and graphs. The tractability of this nonlinear formulation arises from polytopes induced by submodular set functions. Submodular functions possess two highly desirable properties for our problem: (1) they naturally encode combinatorial structure, via diminishing returns and as combinatorial rank functions; and (2) their geometry leads to efficient algorithms.

The resulting combination of the geometries of transportation and submodularity leads to a problem with rich, favorable polyhedral structure and connections to game theory and saddle point optimization. We leverage this structure to solve the *submodular optimal transport* problem via a saddle-point mirror prox algorithm involving alternating projections onto the polytope defined by the transportation constraints and the base polytope associated with the submodular cost function. The former can be done efficiently through Sinkhorn iterations, while the latter, as we will see, can be solved exactly in $O(n \log n)$ time for a suitable class of submodular functions.

Via various applications and experiments, we explore the characteristics of the solutions to this novel transportation problem and demonstrate the efficiency of our algorithms. We show how different submodular functions yield solutions that interpolate between strictly structure-aware transportation plans and structure-agnostic regularized versions of the problem. Besides these synthetic experiments, we evaluate our framework in two real-life applications: domain adaption for digit classification and sentence similarity prediction. In both cases, introducing structure leads to better empirical results.

Contributions. In short, we make the following contributions: (1) we propose a framework for including structured information into optimal transport that integrates concepts from combinatorics to geometry; (2) we show efficient optimization methods that carefully exploit the underlying geometric structure; (3) we demonstrate the utility of this new framework via example applications in domain adaptation and sentence similarity, where our structured couplings outperform classical and class-regularized versions of optimal transport.

2 Background

2.1 Optimal Transport

The original formulation of optimal transport by Gaspar Monge considers two probability measures μ, ν over metric spaces $\mathcal{X}, \mathcal{Y}$, and a measurable cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, which represents the cost of transporting a unit of mass from $x \in \mathcal{X}$ to $y \in \mathcal{Y}$. The problem asks to find a transport map $T : \mathcal{X} \to \mathcal{Y}$ that realizes

$$\inf_T \left\{ \int_{\mathcal{X}} c(x, T(x)) d\mu(x) \mid T_{\#}\mu = \nu \right\}, \tag{1}$$

where $T_{\#}$ denotes the push-forward of μ by T. The solution to (1) might not exist. However, a convex relaxation of the problem due to Kantorovich is guaranteed to have a solution:

$$\inf_\gamma \left\{ \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\gamma(x, y) \mid \gamma \in \Gamma(\mu, \nu) \right\}, \tag{2}$$

where $\Gamma(\mu, \nu)$ is the set of *transportation plans*, i.e., joint distributions with marginals μ and ν. If μ and ν are only available through discrete samples $U := \{\mathbf{x}_i^s\}_{i=1}^n$ and $V := \{\mathbf{x}_i^t\}_{i=1}^m$, the empirical distributions can be written as

$$\mu = \sum_{i=1}^n p_i^s \delta_{\mathbf{x}_i^s}, \quad \nu = \sum_{i=1}^m p_i^t \delta_{\mathbf{x}_i^t} \tag{3}$$

where p_i^s, p_i^t are the probabilities associated with the samples. It is easy to adapt Kantorovich's formulation to this discrete setting. In this case, the space of transportation plans is a polytope:

$$\mathcal{M}_{\mu,\nu} = \{\gamma \in \mathbb{R}_+^{n\times m} \mid \gamma\mathbf{1} = \mu,\ \gamma^T\mathbf{1} = \nu\}. \tag{4}$$

The cost function only needs to be specified for every pair $(\mathbf{x}_i^s, \mathbf{x}_j^t)$, i.e., it is a matrix $C \in \mathbb{R}^{n\times m}$, and the total cost incurred by γ is $\langle \gamma, C\rangle := \sum_{ij} \gamma_{ij} c_{ij}$. Thus, the discrete optimal transport (DOT) problem consists of finding a plan γ that solves

$$\min_{\gamma\in\mathcal{M}_{\mu,\nu}} \langle \gamma, C\rangle. \tag{5}$$

If $n = m$, and μ and ν are uniform measures, $\mathcal{M}_{\mu,\nu}$ is the Birkhoff polytope of size n, and the solutions of (5), which lie in the corners of this polytope, are permutation matrices.

Discrete optimal transport is a linear program, and thus can be solved exactly in $O(n^3 \log n)$ with interior point methods. In practice, a version with entropic smoothing has proven more efficient (Cuturi, 2013):

$$\min_{\gamma\in\mathcal{M}} \langle \gamma, C\rangle - \frac{1}{\lambda} H(\gamma). \tag{6}$$

The solution of this strictly convex optimization problem has the form $\gamma^* = \text{diag}(u)\mathbf{K}\text{diag}(u)$, with $\mathbf{K} = e^{-\frac{C}{\lambda}}$ (entrywise), and can be obtained efficiently via the Sinkhorn-Knopp algorithm, an iterative matrix-scaling procedure (Cuturi, 2013). Besides significant speedups, the smoothed problem often leads to better empirical results in downstream applications.

2.2 Submodularity

A set function $F : 2^V \to \mathbb{R}$ over a ground set V of items is called *submodular* if it satisfies *diminishing returns*: for all $S \subseteq T \subseteq V$ and all v in $V \setminus T$, it holds that

$$F(S \cup \{v\}) - F(S) \geq F(T \cup \{v\}) - F(T) \quad (7)$$

F is called supermodular if $-F$ is submodular, and modular if it is both sub- and supermodular. The tractability of submodular functions arises from the polytopes they define. The *base polytope* of F is

$$\mathcal{B}_F = \{y \in \mathbb{R}^{|V|} \mid y(V) = F(V);\ y(S) \leq F(S)\ \forall S \subseteq V\}.$$

Base polytopes generalize matroid polytopes (convex hulls of combinatorial "independent sets"), and lead to strong links with convexity. The *Lovász extension* of a set function F extends its domain from 2^V to $\mathbb{R}^n_+$ (Lovász, 1982). For any $w \in \mathbb{R}^n_+$, order its coordinates so that $w_1 \geq \cdots \geq w_n$ and define $w_{n+1} = 0$ and $S_j = \{i \mid w_i \geq w_j\}$. The Lovász extension f of F is

$$f(w) = \sum_{j=1}^{n} (w_j - w_{j+1}) F(S_j). \quad (8)$$

If F is submodular, the Lovász extension is equivalent to the support function

$$f(w) = \max_{x \in \mathcal{B}_F} w^T x, \quad (9)$$

which is convex. In fact, f is convex if and only if F is submodular (Lovász, 1982).

3 Optimal Transport with Submodular Costs

In the classical formulation of optimal transport (5), the cost function $\langle \gamma, C \rangle$ is linear in the decision variables γ. This means each potential pairwise assignment γ_{ij} (i.e., every pair (μ_i, ν_j)) is treated independently. But, in some applications, it is desirable to bias certain points to be mapped *together*, i.e., to introduce dependencies between assignments. In our running example of domain adaptation, we want points from the same class to be transported "together". Intuitively, the joint cost of mapping points from the same class to close-by target points should be lower than splitting them apart, even if the transportation distances are the same.

More generally, we might want to encourage mappings of subspaces to subspaces, or, on the contrary, discourage some combinations of assignments. A flexible framework to express such interactions over discrete choices is via submodular functions (Lin and Bilmes, 2011; Jegelka and Bilmes, 2011; Kohli, Osokin, and Jegelka, 2013). Intuitively, property (7) implies that the marginal cost of an additional element decreases as more "compatible" items have already been chosen, and thus it is relatively *cheaper* to select compatible items together (e.g., items from the same group) than non-compatible ones.

To see how submodularity can be leveraged for optimal transport, consider for a moment Monge's formulation (1), where we seek a matching of the elements in U and V with minimal cost. Any matching can be expressed as a set of edges $S = \{(u_1, v_1), \ldots, (u_k, v_k)\}$, and its cost as a set function $F : 2^{|U| \times |V|} \to \mathbb{R}^+$. Under this formulation, the classic definition of optimal transport uses a *modular* cost function:

$$F(S) = \sum_{(u,v) \in S} c_{uv},$$

so the cost of the additional match (u, v) is the same, namely c_{uv}, regardless of what assignments have already been made. If we let F be submodular instead, property (7) implies that the marginal cost of additional edges decreases as the set of matches grows. The magnitude of decrease depends on S, the new item v, and the choice of F. We will channel this decrease to occur only when the additional "item" (assignment (u, v)) is compatible with already chosen "items".

3.1 Submodular cost functions

The rich class of submodular functions allows various types of structural information (compatibility) to be encoded in the cost function. As an example, recall the local consistency structure induced by class labels in domain adaptation. We may divide the support of the source and target distributions μ and ν into regions (subsets of samples) $U_k \subset U$ and $V_l \subset V$. These induce a partition of the set of assignments too:

$$E_{kl} := \{(u, v) \mid u \in U_k,\ v \in V_l\}.$$

Now define

$$F(S) := \sum_{kl} F_{kl}(S \cap E_{kl}), \quad (10)$$

where each F_{kl} is submodular with reduced support E_{kl}. One possible choice for F_{kl} is

$$F_{kl}(S) = g_{kl}\Big(\sum_{(u,v) \in S \cap E_{kl}} C_{uv} \Big), \quad (11)$$

where $C_{ij} \in \mathbb{R}^+$ is the ground metric cost between x_i^s and x_j^t, and $g_{kl} : \mathbb{R} \to \mathbb{R}$ are scalar monotone increasing concave functions whose effect is to dampen the cost of additional edges between the partitions U_l and V_k, thus encouraging edge selections that map most of the mass in U_l to the same V_k. To grant discounts only after a sufficient number of assignments have been chosen from a group, we may use an explicit threshold, e.g.,

$$g_{kl}(x) = \min\{x, \alpha\} + \sqrt{[x-\alpha]_+}. \quad (12)$$

We use such functions in the clustered point matching, domain adaptation and sentence similarity experiments in Section 5. We may also use subspaces for encoding structure. For example, a smoother grouping of assignments (u, v) could be encoded by stacking feature vectors for u and v into one vector $\phi(u, v)$ and taking $F(S) = \mathrm{rank}(\Phi_S)$, i.e., the rank of the matrix of features of the selected assignments, or the volume $F(S) = \log\det(\Phi_S^\top \Phi_S)$. This function captures discrete groups if the feature vectors are indicator vectors of groups. Other important examples include hierarchical structures and coverage functions.

3.2 Problem Formulation: Submodular optimal transport

The functions defined above have discrete domains, i.e., they correspond to discrete matchings, but we really seek a formulation like (5), with continuous, fractional assignments. The key to obtaining a nonlinear, structured analog of Kantorovich's formulation (2) of the classical problem is the convex *Lovász extension* f of the submodular function F. The above intuitions and effects carry over, and we define the *submodular optimal transport* problem as

$$\min_{\gamma \in \mathcal{M}} f(\gamma) \quad \equiv \quad \min_{\gamma \in \mathcal{M}} \max_{\kappa \in \mathcal{B}_F} \langle \gamma, \kappa \rangle. \quad (13)$$

The right hand side follows since the Lovász extension is also the support function of the submodular base polytope. This relaxation has another advantage: while the discrete version is hard to even solve approximately (Goel et al., 2009), problem (13) is a convex optimization problem on γ.

The new structured optimal transport problem recovers many desirable properties of the original optimal transport formulation. For example, the "distance" implied by it is a semi-metric under mild assumptions (proof in the Supplementary Material):

Lemma 3.1. *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.*

Problem (13) suggests two possible approaches for computing the optimal transport plan γ^*. The left-hand side is a non-smooth but convex optimization problem, which can be solved via subgradient methods. Alternatively, the minimax form is a *smooth* convex-concave optimization over nonempty, closed and convex sets.[2]

[2] $\mathcal{M}, \mathcal{B}_F$, being polytopes, are closed and convex. $\mathcal{M}$ is always nonempty ($\mu\nu^T \in \mathcal{M}$), and so is $\mathcal{B}_F$ (Bach, 2013).

Therefore, (13) is a convex-concave saddle-point problem (Juditsky and Nemirovski, 2011a). The solutions $z^* := (\gamma^*, \kappa^*)$ of this problem, i.e., the *saddle points* $\phi := \langle\cdot,\cdot\rangle$ in $\mathcal{Z} := \mathcal{M} \times \mathcal{B}_F$, satisfy

$$\phi(\gamma^*, \kappa) \le \phi(\gamma^*, \kappa^*) \le \phi(\gamma, \kappa^*) \quad \forall \gamma \in \mathcal{M}, \kappa \in \mathcal{B}_F$$

This formulation gives rise to a primal-dual pair of convex optimization problems:

$$\mathrm{Opt}(P) = \min_{\gamma\in\mathcal{M}} \bar{\phi}(\gamma), \quad \bar{\phi}(\gamma) := \sup_{\kappa\in\mathcal{B}_F} \phi(\gamma, \kappa) \quad (14)$$

$$\mathrm{Opt}(D) = \max_{\kappa\in\mathcal{B}_F} \underline{\phi}(\kappa), \quad \underline{\phi}(\kappa) := \sup_{\gamma\in\mathcal{M}} \phi(\gamma, \kappa) \quad (15)$$

If a saddle point (γ^*, κ^*) exists, then it is a primal-dual optimal pair and $\mathrm{Opt}(P) = \mathrm{Opt}(D)$. Hence, the *saddle-point gap* quantifies the accuracy of a candidate solution $(\hat{\gamma}, \hat{\kappa})$:

$$\begin{aligned}\epsilon_{sp} &= \sup_{\gamma} \phi(\gamma, \hat{\kappa}) - \inf_{\kappa} \phi(\hat{\gamma}, \kappa) \\ &= [\bar{\phi}(\gamma) - \mathrm{Opt}(P)] - [\mathrm{Opt}(D) - \underline{\phi}(\kappa)]\end{aligned}$$

Since ϕ is continuous and convex-concave, and $\mathcal{M}, \mathcal{B}_F$ are convex and bounded, a solution always exists.

Although more involved than the alternative convex optimization approach, this saddle-point formulation results in a smooth objective, which allows for the use of methods with $O(\frac{1}{t})$ convergence rate instead of $O(\frac{1}{\sqrt{t}})$. This, however, comes at the price of a higher cost per iteration. We analyze these opposing effects theoretically in the next section and empirically in Section 5. Beyond these computational issues, the saddle-point formulation provides interesting interpretations of the structured optimal transport problem through the lens of minimax optimization and its well-known connections to game theory and robust optimization.

Game Theoretic Interpretation. The minimax formulation (13) is a *min-max strategy polytope* (MSP) game (Gupta, Goemans, and Jaillet, 2016): a two-player zero-sum game with strategies played over polytopes with payoff function $\langle\gamma, \kappa\rangle$. In this optimal transport game, Player A (the *minimizer*) chooses a transport plan γ between μ and ν, and Player B (the *adversary*) chooses a cost matrix κ from the set of *admissible* costs, i.e., those that lie on the base polytope defined by the submodular cost function F. After this, Player A pays $\langle\gamma, \kappa\rangle$ to Player B. Since the game is guaranteed to have a Nash equilibrium, there is a pair of transport plan γ^* and cost matrix κ^* such that γ^* is optimal for fixed cost κ^* and vice versa.

The shape and size of the adversary's strategy polytope $\mathcal{B}_F$, an $nm - 1$ dimensional set in $\mathbb{R}^{n\times m}$, depends on the characteristics of F. The "more submodular" this function is—i.e., the earlier and sharper the marginal costs decrease—the larger is $\mathcal{B}_F$. If F is modular, the base polytope collapses to a single point, that is, Player B plays a fixed strategy: a ground cost matrix C. The problem then reduces to $\min_{\gamma \in \mathcal{M}} \langle \gamma, C \rangle$: the traditional optimal transport problem (5).

Robust Optimization Interpretation. Problem (13) can also be viewed in the light of *robust optimization* (Ben-Tal, Ghaoui, and Nemirovski, 2009; Bertsimas, Brown, and Caramanis, 2011), where uncertain observations are treated in a worst-case scenario. Structured optimal transport could then be viewed as a transportation problem with uncertain cost matrix κ, where we aim for a solution that is robust to any fluctuation of costs within the confidence set $\mathcal{B}_F$.

3.3 Further related work

Courty et al. (2017) propose to include structural information into the standard transportation cost by adding a group-norm regularizer. In contrast, our polyhedral approach directly modifies the linear cost function, does not need a regularization coefficient, allows to integrate a wide set of combinatorial functions, and directly leads to the saddle point connections. Our framework is also fundamentally different from known connections between multi-marginal optimal transport and submodularity (Bach, 2015; Carlier, 2003; Pass, 2015); while that setting is separable over assignments γ_{ij}, the submodularity ranges across assignment pairs between two distributions.

4 Solving the Optimization Problem

4.1 A case for proximal methods

Most popular first-order optimization methods for constrained convex problems fall into two categories: conditional gradient and proximal methods. Methods in the former class, like the Frank-Wolfe algorithm, require solving linear minimization oracles (LMO) as a subroutine. In the case of (13), this means solving a classic (non-regularized) optimal transport problem in each iteration, which is expensive.

On the other hand, proximal methods require mirror map computations and projections. The choice of mirror map is crucial for the efficiency of these methods, and should take into account the geometry of the constraint set. Only if the resulting projections can be easily computed are proximal methods an attractive alternative. As we show below, for appropriately chosen mirror maps this is the case for both constraint sets in problem (13). We briefly discuss all required subroutines in the next section, and present outer optimization algorithms in Section 4.3. We outline the main concepts here; detailed derivations may be found in the Supplementary Material.

4.2 Subroutines: projections and subgradients

Subgradients of f. The subdifferential of f is

$$\partial f(\gamma) = \operatorname*{argmax}_{\kappa \in \mathcal{B}_F} \langle \kappa, \gamma \rangle.$$

Thus, a subgradient of f is computed by a linear optimization over the base polytope, which, despite exponentially many constraints, can be solved by a simple sort via Edmonds' greedy algorithm in $O(N \log N)$ time, where $N = n \times m$ is the dimension of γ.

Projections on the coupling polytope. If we use (negative) entropy as the mirror map in $\mathcal{M}$, i.e., $\Phi_{\mathcal{M}}(\gamma) := H(\gamma) = \sum_{i,j} \gamma_{ij} \ln(\gamma_{ij})$, the projection of a point w onto $\mathcal{M}$ is given by the KL-divergence:

$$\hat{\gamma} = \operatorname*{argmin}_{\gamma \in \mathcal{M}} \mathrm{KL}(\gamma \parallel w). \tag{16}$$

This problem is efficiently solved by the Sinkhorn-Knopp algorithm (Cuturi, 2013; Benamou et al., 2015). An ϵ-accurate solution can be computed in $O(N \log N \epsilon^{-3})$ time (Altschuler, Weed, and Rigollet, 2017), but often much faster empirically (Cuturi, 2013).

Projections on the base polytope. If we use $\Phi_{\mathcal{B}_F}(\kappa) = \frac{1}{2}\|\kappa\|^2$, the resulting Euclidean projection[3] on the base polytope,

$$\hat{\kappa} = \operatorname*{argmin}_{\kappa \in \mathcal{B}_F} \|\kappa - w\|_2^2 \;=\; \operatorname*{argmin}_{\kappa' \in \mathcal{B}_{F-w}} \|\kappa'\|_2^2 + w, \tag{17}$$

is equivalent to minimizing the "shifted" submodular function $F(S) - \sum_{i \in S} w_i$ and can be computed, for instance, via the Fujishige-Wolfe minimum norm point (MNP) algorithm (Wolfe, 1976; Fujishige, Hayashi, and Isotani, 2006), via parametric submodular minimization and with recent cutting-plane algorithms (Lee, Sidford, and Wong, 2015). These generic methods are nevertheless computationally very expensive, except for small problems. But most of the functions of interest, such as the group functions defined in Section 3.1, have additional structure: they are of the form $F(S) = \sum_{i=1}^k F_i(S)$ (also called *decomposable*), each F_i with small support or "simple" structure. Here, "simple" means that the minimum norm point problem can be solved fast. For the functions defined in (11), and more

[3] Perhaps surprisingly, the projection onto the base polytope resulting from choosing $\Phi_{\mathcal{B}_F}(\kappa) := H(\kappa)$ instead is also solved by (17) (Djolonga and Krause, 2015), and hence we may implement mirror descent with either projection.

generally, for certain hierarchical functions (Hochbaum and Hong, 1995; Iwata and Zuiki, 2004), coverage functions (Stobbe and Krause, 2010) and graph cuts on lines (equivalent to Total Variation), this can be solved in $O(m \log m)$ time, where m is the support size of the respective F_i. We provide an $O(m \log m)$ algorithm for our cluster functions in the Supplement. If the supports of the F_i's are disjoint, then the base polytope is a product of polytopes $\mathcal{B}_{F_i}$, and the projection can be computed for each $\mathcal{B}_{F_i}$ separately in parallel. If the supports overlap, then we can still exploit decomposition structure via randomized coordinate descent (Ene, Nguyen, and Végh, 2017), operator splitting methods (Jegelka, Bach, and Sra, 2013; Nishihara, Jegelka, and Jordan, 2014) or others (Stobbe and Krause, 2010) for fast optimization.

4.3 Optimization Algorithms

4.3.1 Convex formulation

We can solve the left hand side of (13) using mirror descent (MDA), shown as Algorithm 1. The choice of entropy mirror map $\Phi(\gamma) = H(\gamma)$ means that every iteration will require a KL-projection onto the base polytope and a subgradient computation, bringing the total cost per iteration to $O(N \log N + N(\log N)\epsilon^{-3})$. For a non-smooth, not strongly convex function like the Lovász extension, MDA converges with rate $O(\frac{1}{\sqrt{t}})$.

4.3.2 Saddle-point formulation

We solve the minimax formulation of problem (13) via either saddle-point mirror-descent (SP-MD) or saddle-point mirror-prox (SP-MP) (Juditsky and Nemirovski, 2011a; Juditsky and Nemirovski, 2011b). The latter enjoys a faster convergence rate, at the cost of doubling the per-iteration cost, requiring two projections onto each of $\mathcal{M}$ and $\mathcal{B}_F$. In either case, the setup is as follows. Let $\Phi_{\mathcal{M}}(\gamma)$ and $\Phi_{\mathcal{B}_F}(\kappa)$ be mirror maps on $\mathcal{M}$ and $\mathcal{B}_F$, then the mirror map for the joint variable $z = (\gamma, \kappa) \in \mathcal{Z} := \mathcal{M} \times \mathcal{B}_F$ is $\Phi(z) = \Phi_{\mathcal{M}}(\gamma) + \Phi_{\mathcal{B}_F}(\kappa)$, and a first-order oracle F for ϕ is required to obtain subgradients in $\partial\phi(z) = \{\partial_\gamma[\phi(\gamma,\kappa)]\} \times \{\partial_\kappa[-\phi(\gamma,\kappa)]\}$. Thus, both the gradient computation and projection decouple over κ and γ, and we can use the projections described in Section 4.2. The final SP-MP method for solving problem (13) is shown as Algorithm 2. The (simpler) SP-MD is analogous with a single Sinkhorn/projection step. Compared to MDA and SP-MD, the mirror prox version enjoys a better convergence rate of $O(\frac{1}{t})$. Using the fast projection method for the cluster-based functions proposed here (Eq. 10), the total cost per iteration in either SP-MD and SP-MP is $O(N(\log N)\epsilon^{-3} + K \log K)$, where K is the size of the largest cluster.

Algorithm 1 MDA for Structured Optimal Transport

Input: Initial point γ_0 and initial step size η_0
while $\epsilon < tol$ **do**
  $g_t \leftarrow$ EDMONDS(f, γ_t)
  $\tilde{\gamma}_{t+1} \leftarrow$ SINKHORN$(\gamma_t \circ \exp\{-\eta_t g_t\})$
  $\gamma_{t+1} \leftarrow [\sum_{s=1}^{t+1} \eta_s]^{-1} \sum_{s=1}^{t+1} \eta_s \tilde{\gamma}_s$
  $\epsilon \leftarrow f(\gamma_t) - f(\gamma_{t+1})$
  $t \leftarrow t+1$
end while

Algorithm 2 Saddle Point Mirror Prox for Structured Optimal Transport

Input: Initial point $z^0 = (\gamma_0, \kappa_0)$ and step size η_0
while $\epsilon_{SP} < tol$ **do**
  // Mirror step on true gradient
  $u_{t+1} \leftarrow$ SINKHORN$(\gamma_t \circ \exp\{-\eta_t \kappa_t\})$
  $v_{t+1} \leftarrow$ BASEPOLYPROJECT$(\kappa_t + \eta_t \gamma_t)$
  // Mirror step on proxy gradient
  $\gamma_{t+1} \leftarrow$ SINKHORN$(\gamma_t \circ \exp\{-\eta_t v_{t+1}\})$
  $\kappa_{t+1} \leftarrow$ BASEPOLYPROJECT$(kappa_t + \eta_t u_{t+1})$
  // Compute saddle point gap of current solution
  $z^{t+1} \leftarrow [\sum_{s=1}^{t+1} \eta_s]^{-1} \sum_{s=1}^{t+1} \eta_s(\gamma_s, \kappa_s)$
  $\epsilon_{SP} \leftarrow$ SADDLEGAP(z^t)
  $t \leftarrow t+1$
end while

Initialization A simple choice for γ_0 is $\mu\nu^T$. For κ_0, a random corner in the base polytope[4] can be used, however, we found that initializing it as the projection of C onto $\mathcal{B}_F$ often results in faster convergence.

5 Experimental Results

Our implementation of Algorithms 1 and 2 uses the Python Optimal Transport library (Flamary and Courty, 2017) for entropic projections onto the transport polytope. For the projections onto the base polytope required by SP-MP (Alg. 2), we use a tailored algorithm for decomposable functions (detailed in the Supplementary Material) and RCDM (Ene and Nguyen, 2015) when the supports are not disjoint. All experiments were run on a 2.8GHz Intel Core i7 Processor.

5.1 Clustered Point Cloud Matching

Synthetic Point Clouds. In our first set of experiments, we seek to understand the characteristics of the transport plans obtained with our structured optimal transport (SOT) framework. For this, we generate two point clouds in $\mathbb{R}^2$ from two distinct 3-gaussian mixture distributions (20 points each, 60/20/20% class splits). We use the class labels to define a sum-of-clusters func-

[4] Computed, e.g., by evaluating f for random $w \in \mathbb{R}^{n \times m}$.

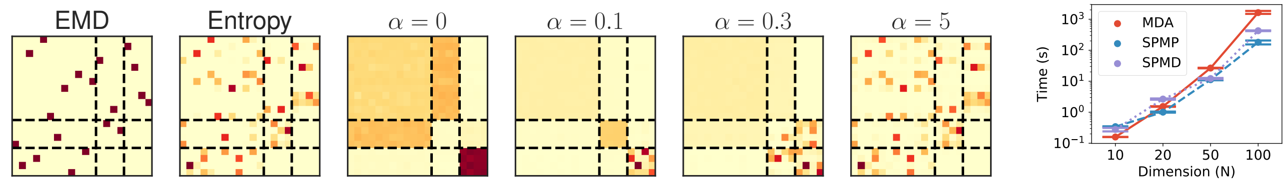

Figure 1: Optimal transport plans for clustered point matching obtained with two structure-agnostic formulations (EMD, entropy-regularized) and our submodular approach with varying concavity threshold parameter α (Eqn. (12)). Dashed lines show class partitions. **Right:** Runtimes for alternative optimization methods.

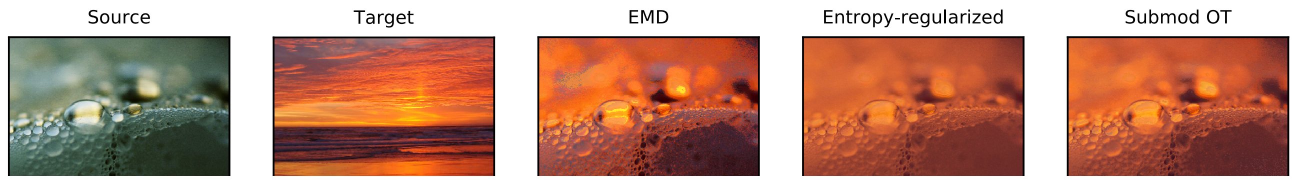

Figure 2: Color transfer with various optimal transport methods. The pixels in the source image get their color from the transported pixels in the target image.

tion as in (11), using square-root thresholding functions (12) for varying values of α. The optimal coupling matrices are shown in Figure 1. As expected, lower values of α enforce cluster structure more aggressively, while for larger α the cost effectively becomes modular, causing the solution to resemble those of the unstructured OT formulations. In terms of empirical runtimes (Fig. 1, right), SP-MP generally outperforms both SP-MD and MDA except in the very low sample size regime.

Color transfer. An interesting application of this matching with group information is color transfer. Here, we seek to transfer the colors of one image (the *target* color scheme) into another one, the *source*. To do so, we view pixels as points in RGB space, transport them using optimal transport, and assign their color to the matched pixels. Here we define partitions through superpixels obtained by segmentation (Felzenszwalb and Huttenlocher, 2004). The example in Figure 2 shows that including structure in the cost function results in a coloring scheme that is more uniform that the EMD variant and sharper than the entropy-regularized one.

5.2 Domain Adaptation

Domain adaptation can be naturally cast as a transportation problem. When modeling the source and target distributions via discrete samples, DOT yields an optimal transport plan γ^* between the two samples, according to which source points can be "transported" to the target domain through the *barycentric* mapping implicitly defined by γ^* (Villani, 2008, Chapter 7).

In our motivating example of domain adaptation for classification, we wish to incorporate any available class labels on either domain into the cost function, so as to encourage points of the same class to be mapped to the same region of the target space. This is seamlessly attainable with our proposed framework and the cluster functions defined before (11). In the experiments below, we partition the source samples according to their class label, but we do not use the target labels (i.e., every target sample forms its own cluster), so as to simulate the harder—and more realistic—unsupervised domain adaptation setting.

We test this adaptation approach on the benchmark USPS and MNIST digit classification datasets. We preprocess the data by normalizing, and downscale MNIST to the 16×16 size of USPS. Here, we simulate an extreme adaptation setting where only 100 samples of each domain are provided, and no target labels are available. We train a 1-NN classifier on the transported samples, and use it to predict labels on the test set (10K examples for MNIST, ~2K for USPS).

We compare our method (using (11) with (12), and a default $\alpha = 0.2$ threshold) against the two class-regularized OT formulations of Courty et al. (2017): one using an ℓ_p–ℓ_1 group-sparsity norm, and the other a Laplacian regularization term. We also compare against the original and entropy-regularized formulations, neither of which uses class labels. The results in Table 1 show that the submodular formulation achieves better accuracy in both directions of adaptation, and exhibits much clearer block-diagonal structure in the coupling matrix (Figure 3). We emphasize that the target labels are not used when defining the groupings of the submodular function, so this block structure is obtained solely by encouraging source points with the same label to be mapped together. Example source and transported digits are shown in the Supplement.

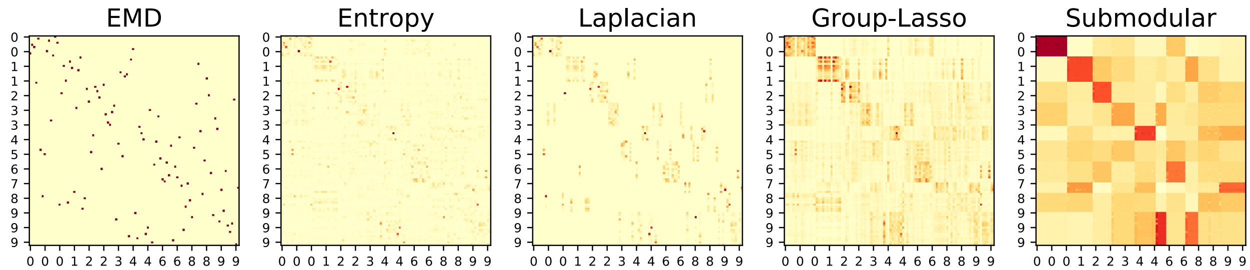

Figure 3: Optimal transport plans for the MNIST→USPS adaptation task. Rows and columns are sorted by class.

Method	MNIST→USPS	USPS→MNIST
No adaption	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
Struct-OT	**62.97**	**58.34**

Table 1: Results on digit recognition adaptation. Values shown are prediction accuracy (%).

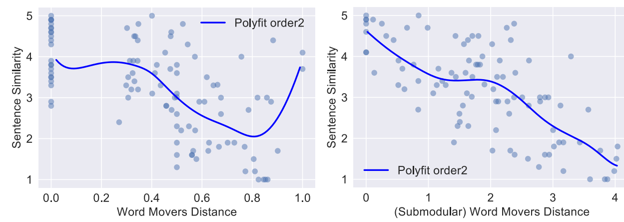

Figure 4: Sentence similarity prediction with two classes of optimal transport distances over sentences.

5.3 Syntax-aware Word Mover's Distance

The *Word Mover's Distance* (WMD) is an application of optimal transport to natural language processing (Kusner et al., 2015). It measures dissimilarity between strings (sentences or documents) by computing the cost of "moving" the words from one to the other, using a ground metric of distances between vector-space embeddings of words. The WMD, however, is syntax-agnostic, i.e., it does not take into account word ordering. That is, the cost of "moving" a word u_i in sentence U to v_j in sentence V depends only on their distance in the embedded space, and not on their relative positions in the two sentences. When using WMD to predict sentence similarity of long sentences with subclauses, this approach can have obvious drawbacks, like transporting words across noun-phrase boundaries.

We can obtain a syntax-aware alternative to WMD with a simple clustered cost function as before, where now each n-gram in a sentence defines a group (i.e., we allow overlaps between the groups). With this, we are encouraging neighboring words in a sentence to be matched to neighboring words in the other. Word-to-word costs are defined as before. We compare this distance against the original WMD in a simple sentence similarity task: the SICK dataset, consisting of pairs of English sentences labeled with human-generated similarity scores. We randomly select 100 sentences with at most 10 words from the train and test folds, we compute optimal transport distances between all training pairs, and then fit a non-parametric regression model to predict similarity scores from these distances. At test time, given a pair of sentences, we compute the distance between them and use the regression model to predict their similarity. The distances, gold similarity scores and fitted models are shown in Figure 4. The WMD model obtains a mean squared error of 0.67 (Spearman's ρ of .71), while our proposed syntax-aware version has a much better correlation with gold similarity scores (MSE=0.59, $\rho = .75$).

6 Discussion

We proposed a generic framework for including structural information into optimal transport problems, which are finding a growing range of applications in machine learning. While we demonstrated the utility of the framework via examples in domain adaptation, color transfer and sentence similarity, our framework can encode a variety of structures beyond these settings, since it allows arbitrary submodular functions. This choice will depend on the specifics of the problem and the efficiency with which the projections can be solved. The overall resulting convex optimization problem is efficiently solvable via mirror descent methods. For very large problems or general submodular functions, approximate or stochastic submodular optimization subroutines (if applicable) may be suitable.

In fact, the flexibility of our framework goes beyond submodularity; any convex function with bounded closed gradient maps would work as f. Here, we explicitly chose submodular functions due to their favorable geometry and resulting tractability, and their ability to encode a wide range of combinatorial structures.

Acknowledgements

This research was partially supported by an NSF CAREER award 1553284 and a CONACYT graduate fellowship. The authors would like to thank Suvrit Sra for planting the seed for this work by asking whether structure optimal transport is possible.

References

Altschuler, Jason, Jonathan Weed, and Philippe Rigollet (2017). "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration". In: *Advances in Neural Information Processing Systems*, pp. 1961–1971. arXiv: 1705.09634.

Arjovsky, Martin, Soumith Chintala, and Léon Bottou (2017). "Wasserstein Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. PMLR, pp. 214–223.

Bach, Francis (2013). "Learning with submodular functions: A convex optimization perspective". In: *Foundations and Trends in Machine Learning* 6.2-3, pp. 145–373.

Bach, Francis (2015). "Submodular Functions: from Discrete to Continous Domains". In: *CoRR* abs/1511.0.

Ben-Tal, Aharon, Laurent El Ghaoui, and Arkadi Nemirovski (2009). *Robust optimization*, p. 542. ISBN: 9780691143682. arXiv: 1011.1669v3.

Benamou, Jean-David et al. (2015). "Iterative Bregman projections for regularized transportation problems". In: *SIAM Journal on Scientific Computing* 37.2, A1111–A1138.

Bertsimas, Dimitris, David B. Brown, and Constantine Caramanis (2011). "Theory and Applications of Robust Optimization". In: *SIAM Review* 53.3, pp. 464–501. ISSN: 0036-1445. arXiv: 1010.5445.

Carlier, Guillaume (2003). "On a class of multidimensional optimal transportation problems". In: *Journal of convex analysis* 10.2, pp. 517–530.

Courty, Nicolas et al. (2017). "Optimal Transport for Domain Adaptation". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* X.X, pp. 1–14. ISSN: 0162-8828. arXiv: 1507.00504.

Cuturi, Marco (2013). "Sinkhorn distances: Lightspeed computation of optimal transport". In: *Advances in Neural Information Processing Systems*, pp. 1–9. arXiv: 1306.0895v1.

Djolonga, Josip and Andreas Krause (2015). "Scalable Variational Inference in Log-supermodular Models". In: *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 37. PMLR, pp. 1804–1813. arXiv: 1502.06531v2.

Ene, Alina and Huy L. Nguyen (2015). "Random Coordinate Descent Methods for Minimizing Decomposable Submodular Functions". In: *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, pp. 787–795. arXiv: 1502.02643v1.

Ene, Alina, Huy L. Nguyen, and László A. Végh (2017). "Decomposable Submodular Function Minimization: Discrete and Continuous". In: *Advances in Neural Information Processing Systems 30*, pp. 2870–2880. arXiv: 1703.01830.

Felzenszwalb, Pedro F and Daniel P Huttenlocher (2004). "Efficient graph-based image segmentation". In: *International Journal of Computer Vision* 59.2, pp. 167–181.

Flamary, Rémi and Nicolas Courty (2017). "POT: Python Optimal Transport library". In:

Frogner, Charlie et al. (2015). "Learning with a Wasserstein Loss". In: *Neural Information Processing Systems*, pp. 2053–2061. arXiv: 1506.05439.

Fujishige, Satoru, Takumi Hayashi, and Shigueo Isotani (2006). "The Minimum-Norm-Point Algorithm Applied to Submodular Function Minimization and Linear Programming". In: *RIMS preprint 1571*, pp. 1–19.

Galichon, Alfred (2016). *Optimal Transport Methods in Economics*. Princeton University Press.

Gangbo, W. and R.J. McCann (2000). "Shape recognition via Wasserstein distance". In: *Quarterly of Applied Mathematics* 58.4, pp. 705–738. ISSN: 0033-569x.

Genevay, Aude et al. (2016). "Stochastic optimization for large-scale optimal transport". In: *Advances in Neural Information Processing Systems*, pp. 3432–3440.

Glimm, T and V Oliker (2003). "Optical Design of Single Reflector Systems and the Monge–Kantorovich Mass Transfer Problem". In: *Journal of Mathematical Sciences* 117.3, pp. 4096–4108. ISSN: 1573-8795.

Goel, Gagan et al. (2009). "Approximability of combinatorial problems with multi-agent submodular cost functions". In: *Foundations of Computer Science, 2009*. IEEE, pp. 755–764.

Gupta, Swati, Michel X Goemans, and Patrick Jaillet (2016). "Solving Combinatorial Games using Products, Projections and Lexicographically Optimal Bases". In: *CoRR* abs/1603.0.

Hochbaum, D S and S.-P. Hong (1995). "About strongly polynomial time algorithms for quadratic optimization over submodular constraints". In: *Mathematical Programming*, pp. 269–309.

Iwata, S and N Zuiki (2004). "A network flow approach to cost allocation for rooted trees". In: *Networks* 44, pp. 297–301.

Jegelka, Stefanie, Francis Bach, and Suvrit Sra (2013). "Reflection methods for user-friendly submodular optimization". In: *Advances in Neural Information Processing Systems*, pp. 1313–1321. arXiv: 1311.4296v1.

Jegelka, Stefanie and Jeff Bilmes (2011). "Submodularity beyond submodular energies: Coupling edges in graph cuts". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1897–1904. ISBN: 9781457703942.

Jordan, Richard, David Kinderlehrer, and Felix Otto (1998). "The Variational Formulation of the Fokker–Planck Equation". In: *SIAM J. Math. Anal.* 29.1, pp. 1–17. ISSN: 0036-1410.

Juditsky, Anatoli and Arkadi Nemirovski (2011a). "First order methods for nonsmooth convex large-scale optimization, I: general purpose methods". In: *Optimization For Machine Learning*. Ed. by Suvrit Sra, Sebastian Nowozin, and Stephen J Wright. MIT Press. Chap. 5, pp. 121–148.

Juditsky, Anatoli and Arkadi Nemirovski (2011b). "First order methods for nonsmooth convex large-scale optimization, II: utilizing problem structure". In: *Optimization For Machine Learning*. Ed. by Suvrit Sra, Sebastian Nowozin, and Stephen J Wright. MIT Press. Chap. 6, pp. 149–183.

Kohli, Pushmeet, Anton Osokin, and Stefanie Jegelka (2013). "A Principled Deep Random Field Model for Image Segmentation". In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1971–1978. ISSN: 10636919.

Kusner, Matt J et al. (2015). "From Word Embeddings To Document Distances". In: *Proceedings of The 32nd International Conference on Machine Learning* 37, pp. 957–966.

Lee, Yin Tat, Aaron Sidford, and Sam Chiu-wai Wong (2015). "A faster cutting plane method and its implications for combinatorial and convex optimization". In: *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*. IEEE, pp. 1049–1065.

Lin, Hui and Jeff Bilmes (2011). "Optimal Selection of Limited Vocabulary Speech Corpora". In: *Twelfth Annual Conference of the International Speech Communication Association*, pp. 1489–1492.

Lovász, L (1982). "Mathematical programming – The State of the Art". In: ed. by A. Bachem, M. Grötschel, and B. Korte. Springer-Verlag Berlin Heidelberg. Chap. Submodular, pp. 235–257. ISBN: 978-3-642-68876-8.

Mallows, C. L. (1972). "A Note on Asymptotic Joint Normality". In: *The Annals of Mathematical Statistics* 43.2, pp. 508–515. ISSN: 0003-4851.

Nishihara, Robert, Stefanie Jegelka, and Michael I Jordan (2014). "On the convergence rate of decomposable submodular function minimization". In: *Advances in Neural Information Processing Systems*, pp. 640–648.

Pass, Brendan (2015). "Multi-marginal optimal transport: theory and applications". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 49.6, pp. 1771–1790.

Pimentel, Samuel D et al. (2015). "Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons". In: *Journal of the American Statistical Association* 110.510, pp. 515–527.

Rosenbaum, Paul R and Donald B Rubin (1985). "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score". In: *The American Statistician* 39.1, pp. 33–38.

Solomon, Justin et al. (2015). "Convolutional Wasserstein Distances : Efficient Optimal Transportation on Geometric Domains". In: *ACM Transactions of Graphics (TOG)* 34, p. 66. ISSN: 07300301.

Stobbe, P and A Krause (2010). "Efficient Minimization of Decomposable Submodular Functions". In: *Advances in Neural Information Processing Systems*, pp. 2208–2216.

Villani, Cédric (2008). *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media.

Wolfe, Philip (1976). "Finding the nearest point in A polytope". In: *Mathematical Programming* 11.1, pp. 128–149. ISSN: 00255610.