Smooth and Sparse Optimal Transport

Mathieu Blondel
NTT Communication Science Laboratories

Vivien Seguy*
Kyoto University

Antoine Rolet*
Kyoto University

Abstract

Entropic regularization is quickly emerging as a new standard in optimal transport (OT). It enables to cast the OT computation as a *differentiable* and *unconstrained* convex optimization problem, which can be efficiently solved using the Sinkhorn algorithm. However, entropy keeps the transportation plan strictly positive and therefore completely dense, unlike unregularized OT. This lack of sparsity can be problematic in applications where the transportation plan itself is of interest. In this paper, we explore regularizing the primal and dual OT formulations with a *strongly convex* term, which corresponds to relaxing the dual and primal constraints with *smooth* approximations. We show how to incorporate squared 2-norm and group lasso regularizations within that framework, leading to *sparse* and *group-sparse* transportation plans. On the theoretical side, we bound the approximation error introduced by regularizing the primal and dual formulations. Our results suggest that, for the regularized primal, the approximation error can often be smaller with squared 2-norm than with entropic regularization. We showcase our proposed framework on the task of color transfer.

1 Introduction

Optimal transport (OT) distances (a.k.a. Wasserstein or earth mover's distances) are a powerful computational tool to compare probability distributions and have recently found widespread use in machine learning [Cuturi, 2013, Solomon et al., 2014, Kusner et al., 2015, Courty et al., 2016, Arjovsky et al., 2017]. While OT distances exhibit a unique ability to capture the geometry of the data, their application to large-scale problems has been largely hampered by their high computational cost. Indeed, computing OT distances involves a linear program, which takes super-cubic time in the data size to solve using state-of-the-art network-flow algorithms. Related to the Schrödinger problem [Schrödinger, 1931, Léonard, 2012], entropy-regularized OT distances have recently gained popularity due to their desirable properties [Cuturi, 2013]. Their computation involves a comparatively easier *differentiable* and *unconstrained* convex optimization problem, which can be solved using the Sinkhorn algorithm [Sinkhorn and Knopp, 1967]. Unlike unregularized OT distances, entropy-regularized OT distances are also differentiable w.r.t. their inputs, enabling their use as a loss function in a machine learning pipeline [Frogner et al., 2015, Rolet et al., 2016].

Despite its considerable merits, however, entropy-regularized OT has some limitations, such as introducing blurring in the optimal transportation plan. While this nuisance can be reduced by using small regularization, this requires a carefully engineered implementation, since the naive Sinkhorn algorithm is numerically unstable in this regime [Schmitzer, 2016]. More importantly, the entropy term keeps the transportation plan strictly positive and therefore completely dense, unlike unregularized OT. This lack of sparsity can be problematic when the optimal transportation plan itself is of interest, *e.g.*, in color transfer [Pitié et al., 2007], domain adaptation [Courty et al., 2016] and ecological inference [Muzellec et al., 2017]. Sparsity in these applications is motivated by the principle of parsimony (simple solutions should be preferred) and by the enhanced interpretability of transportation plans.

Our contributions. This background motivates us to study regularization schemes that lead to *smooth* optimization problems (*i.e.*, differentiable everywhere and with Lipschitz continuous gradient) while retaining the desirable property of *sparse* transportation plans. To do so, we make the following contributions.

We regularize the primal with an arbitrary strongly convex term and derive the corresponding smoothed

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84. *Work performed during an internship at NTT.

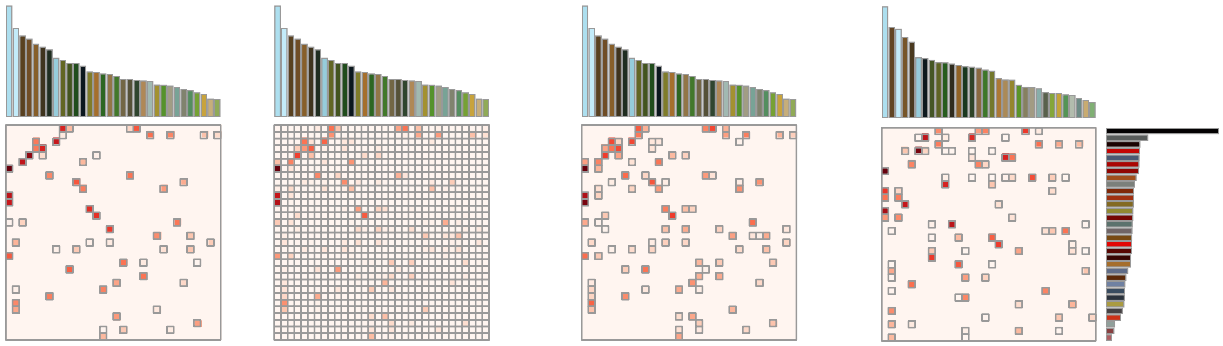

Figure 1: Comparison of transportation plans obtained by different formulations on the application of color transfer. The top and right histograms represent the color distributions $\boldsymbol{a} \in \triangle^m$ and $\boldsymbol{b} \in \triangle^n$ of two images. For the sake of illustration, the number of colors is reduced to $m = n = 32$, using k-means clustering. Small squares indicate non-zero elements in the obtained transportation plan, denoted by T throughout this paper. The sparsity indicated below each graph is the percentage of zero elements in T. The weight of the elements of T indicates the extent to which colors from one image must be transferred to colors from the other image. Like unregularized OT (first from left), but unlike entropy-regularized OT (second from left), our squared 2-norm regularized OT (third from left) is able to produce sparse transportation plans. This is also the case of our relaxed primal (not shown) and semi-relaxed primal (fourth from left) formulations.

dual and semi-dual. Our derivations abstract away regularization-specific terms in an intuitive way (§3). We show how incorporating squared 2-norm and group-lasso regularizations within that framework leads to sparse solutions. This is illustrated in Figure 1 for squared 2-norm regularization.

Next, we explore the opposite direction: replacing one or both of the primal marginal constraints with approximate smooth constraints. When using the squared Euclidean distance to approximate the constraints, we show that this can be interpreted as adding squared 2-norm regularization to the dual (§4). As illustrated in Figure 1, that approach also produces sparse transportation plans.

For both directions, we bound the approximation error caused by regularizing the original OT problem. For the regularized primal, we show that the approximation error of squared 2-norm regularization can be smaller than that of entropic regularization (§5). Finally, we showcase the proposed approaches empirically on the task of color transfer (§6).

An open-source Python implementation is available at `https://github.com/mblondel/smooth-ot`.

Notation. We denote scalars, vectors and matrices using lower-case, bold lower-case and upper-case letters, *e.g.*, t, $\boldsymbol{t}$ and T, respectively. Given a matrix T, we denote its elements by $t_{i,j}$ and its columns by $\boldsymbol{t}_j$. We denote the set $\{1, \ldots, m\}$ by $[m]$. We use $\|\cdot\|_p$ to denote the p-norm. When $p = 2$, we simply write $\|\cdot\|$. We denote the $(m-1)$-dimensional probability simplex by $\triangle^m \coloneqq \{\boldsymbol{y} \in \mathbb{R}_+^m : \|\boldsymbol{y}\|_1 = 1\}$ and the Euclidean projection onto it by $P_{\triangle^m}(\boldsymbol{x}) \coloneqq \operatorname{argmin}_{\boldsymbol{y} \in \triangle^m} \|\boldsymbol{y} - \boldsymbol{x}\|^2$. We denote $[\boldsymbol{x}]_+ \coloneqq \max(\boldsymbol{x}, \boldsymbol{0})$, performed element-wise.

2 Background

Convex analysis. The convex conjugate of a function $f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ is defined by

$$f^*(\boldsymbol{x}) \coloneqq \sup_{\boldsymbol{y} \in \operatorname{dom} f} \boldsymbol{y}^\top \boldsymbol{x} - f(\boldsymbol{y}). \tag{1}$$

If f is strictly convex, then the supremum in (1) is uniquely achieved. Then, from Danskin's theorem (1966), it is equal to the gradient of f^*:

$$\nabla f^*(\boldsymbol{x}) = \operatorname*{argmax}_{\boldsymbol{y} \in \operatorname{dom} f} \boldsymbol{y}^\top \boldsymbol{x} - f(\boldsymbol{y}).$$

The dual of a norm $\|\cdot\|$ is defined by $\|\boldsymbol{x}\|_* \coloneqq \sup_{\|\boldsymbol{y}\| \le 1} \boldsymbol{y}^\top \boldsymbol{x}$. We say that a function is γ-smooth w.r.t. a norm $\|\cdot\|$ if it is differentiable everywhere and its gradient is γ-Lipschitz continuous w.r.t. that norm. Strong convexity plays a crucial role in this paper due to its well-known duality with smoothness: f is γ-strongly convex w.r.t. a norm $\|\cdot\|$ if and only if f^* is $\frac{1}{\gamma}$-smooth w.r.t. $\|\cdot\|_*$ [Kakade et al., 2012].

Optimal transport. We focus throughout this paper on OT between discrete probability distributions $\boldsymbol{a} \in \triangle^m$ and $\boldsymbol{b} \in \triangle^n$. Rather than performing a pointwise comparison of the distributions, OT distances compute the minimal effort, according to some ground cost, for moving the probability mass of one distribution to the other. The modern OT formulation, due to

Kantorovich [1942], is cast as a linear program (LP):

$$\mathrm{OT}(\boldsymbol{a},\boldsymbol{b}) \coloneqq \min_{T\in\mathcal{U}(\boldsymbol{a},\boldsymbol{b})} \langle T, C\rangle, \tag{2}$$

where $\mathcal{U}(\boldsymbol{a},\boldsymbol{b})$ is the transportation polytope

$$\mathcal{U}(\boldsymbol{a},\boldsymbol{b}) \coloneqq \left\{T \in \mathbb{R}_+^{m\times n} \colon T\mathbf{1}_n = \boldsymbol{a}, T^\top \mathbf{1}_m = \boldsymbol{b}\right\}$$

and $C \in \mathbb{R}_+^{m\times n}$ is a cost matrix. The former can be interpreted as the set of all joint probability distributions with marginals $\boldsymbol{a}$ and $\boldsymbol{b}$. Without loss of generality, we will assume $\boldsymbol{a} > 0$ and $\boldsymbol{b} > 0$ throughout this paper (if $a_i = 0$ or $b_j = 0$ then the i^{th} row or the j^{th} column of $T^\star$ is zero). When $n = m$ and C is a distance matrix raised to the power p, $\mathrm{OT}(\cdot,\cdot)^{\frac{1}{p}}$ is a distance on $\triangle^n$, called the Wasserstein distance of order p [Villani, 2003, Theorem 7.3]. The dual LP is

$$\mathrm{OT}(\boldsymbol{a},\boldsymbol{b}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}\in\mathcal{P}(C)} \boldsymbol{\alpha}^\top \boldsymbol{a} + \boldsymbol{\beta}^\top \boldsymbol{b}, \tag{3}$$

where $\mathcal{P}(C) \coloneqq \{\boldsymbol{\alpha} \in \mathbb{R}^m, \boldsymbol{\beta} \in \mathbb{R}^n \colon \alpha_i + \beta_j \leq c_{i,j}\}$. Keeping $\boldsymbol{\alpha}$ fixed, an optimal solution w.r.t. $\boldsymbol{\beta}$ is

$$\beta_j = \min_{i\in[m]} c_{i,j} - \alpha_i, \quad \forall j \in [n],$$

which is the so-called *c-transform*. Plugging it back into the dual, we get the "semi-dual"

$$\mathrm{OT}(\boldsymbol{a},\boldsymbol{b}) = \max_{\boldsymbol{\alpha}\in\mathbb{R}^m} \boldsymbol{\alpha}^\top \boldsymbol{a} - \sum_{j=1}^n b_j \max_{i\in[m]}(\alpha_i - c_{i,j}). \tag{4}$$

For a recent and comprehensive survey of computational OT, see [Peyré and Cuturi, 2017].

3 Strong primal ↔ Relaxed dual

We study in this section adding strongly convex regularization to the primal problem (2). We define

Definition 1 *Strongly convex primal*

$$\mathrm{OT}_\Omega(\boldsymbol{a},\boldsymbol{b}) \coloneqq \min_{T\in\mathcal{U}(\boldsymbol{a},\boldsymbol{b})} \langle T, C\rangle + \sum_{j=1}^n \Omega(\boldsymbol{t}_j), \tag{5}$$

where we assume that Ω *is strongly convex over the intersection of* $\operatorname{dom}\Omega$ *and either* $\mathbb{R}_+^m$ *or* $\triangle^m$.

These assumptions are sufficient for (5) to be strongly convex w.r.t. $T \in \mathcal{U}(\boldsymbol{a},\boldsymbol{b})$. On first sight, solving (5) does not seem easier than (2). As we shall now see, the main benefit occurs when switching to the dual.

3.1 Smooth relaxed dual formulation

Let the (non-smooth) indicator function of the non-positive orthant be defined as

$$\delta(\boldsymbol{x}) \coloneqq \begin{cases} 0, & \text{if } \boldsymbol{x} \leq 0 \\ \infty, & \text{o.w.} \end{cases} = \sup_{\boldsymbol{y}\geq 0} \boldsymbol{y}^\top \boldsymbol{x}.$$

To define a smoothed version of δ, we take the convex conjugate of Ω, restricted to the non-negative orthant:

$$\delta_\Omega(\boldsymbol{x}) \coloneqq \sup_{\boldsymbol{y}\geq 0} \boldsymbol{y}^\top \boldsymbol{x} - \Omega(\boldsymbol{y}). \tag{6}$$

If Ω is γ-strongly convex over $\mathbb{R}_+^m \cap \operatorname{dom}\Omega$, then δ_Ω is $\frac{1}{\gamma}$-smooth and its gradient is $\nabla\delta_\Omega(\boldsymbol{x}) = \boldsymbol{y}^\star$, where $\boldsymbol{y}^\star$ is the supremum of (6). We next show that δ_Ω plays a crucial role in expressing the dual of (5), which is a smooth optimization problem in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

Proposition 1 *Smooth relaxed dual*

$$\mathrm{OT}_\Omega(\boldsymbol{a},\boldsymbol{b}) = \max_{\substack{\boldsymbol{\alpha}\in\mathbb{R}^m\\ \boldsymbol{\beta}\in\mathbb{R}^n}} \boldsymbol{\alpha}^\top \boldsymbol{a} + \boldsymbol{\beta}^\top \boldsymbol{b} - \sum_{j=1}^n \delta_\Omega(\boldsymbol{\alpha} + \beta_j \mathbf{1}_m - \boldsymbol{c}_j) \tag{7}$$

The optimal solution $T^\star$ *of* (5) *can be recovered from* $(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$ *by* $\boldsymbol{t}_j^\star = \nabla\delta_\Omega(\boldsymbol{\alpha}^\star + \beta_j^\star \mathbf{1}_m - \boldsymbol{c}_j) \quad \forall j \in [n]$.

For a proof, see Appendix A.1. Intuitively, the hard dual constraints $\alpha_i + \beta_j - c_{i,j} \leq 0$ $\forall i \in [m]$ $\forall j \in [n]$, which we can write $\sum_{j=1}^n \delta(\boldsymbol{\alpha} + \beta_j \mathbf{1}_m - \boldsymbol{c}_j)$, are now relaxed with *soft* ones by substituting δ with δ_Ω.

3.2 Smoothed semi-dual formulation

We now derive the semi-dual of (5), *i.e.*, the dual (7) with one of the two variables eliminated. Without loss of generality, we proceed to eliminate $\boldsymbol{\beta}$. To do so, we use the notion of smoothed max operator. Notice that

$$\max(\boldsymbol{x}) \coloneqq \max_{i\in[m]} x_i = \sup_{\boldsymbol{y}\in\triangle^m} \boldsymbol{y}^\top \boldsymbol{x} \quad \forall \boldsymbol{x} \in \mathbb{R}^m.$$

This is indeed true, since the supremum is always achieved at one of the simplex vertices. To define a smoothed max operator [Nesterov, 2005], we take the conjugate of Ω, this time restricted to the simplex:

$$\max{}_\Omega(\boldsymbol{x}) \coloneqq \sup_{\boldsymbol{y}\in\triangle^m} \boldsymbol{y}^\top \boldsymbol{x} - \Omega(\boldsymbol{y}). \tag{8}$$

If Ω is γ-strongly convex over $\triangle^m \cap \operatorname{dom}\Omega$, then $\max_\Omega$ is $\frac{1}{\gamma}$-smooth and its gradient is defined by $\nabla\max_\Omega(\boldsymbol{x}) = \boldsymbol{y}^\star$, where $\boldsymbol{y}^\star$ is the supremum of (8). We next show that $\max_\Omega$ plays a crucial role in expressing the conjugate of OT_Ω.

Lemma 1 *Conjugate of* OT_Ω *w.r.t. its first argument*

$$\mathrm{OT}_\Omega^*(\boldsymbol{\alpha},\boldsymbol{b}) = \sup_{\boldsymbol{a}\in\triangle^m} \boldsymbol{\alpha}^\top \boldsymbol{a} - \mathrm{OT}_\Omega(\boldsymbol{a},\boldsymbol{b}) = \sum_{j=1}^n b_j \max{}_{\Omega_j}(\boldsymbol{\alpha} - \boldsymbol{c}_j), \tag{9}$$

where $\Omega_j(\boldsymbol{y}) \coloneqq \frac{1}{b_j}\Omega(b_j \boldsymbol{y})$.

A proof is given in Appendix A.2. With the conjugate, we can now easily express the semi-dual of (5), which involves a smooth optimization problem in $\boldsymbol{\alpha}$.

Table 1: Closed forms for δ_Ω (used in smoothed dual), $\max_{\Omega_j}$ (used in smoothed semi-dual) and their gradients.

	$\Omega(\boldsymbol{y})$	$\delta_\Omega(\boldsymbol{x})$	$\nabla\delta_\Omega(\boldsymbol{x})$	$\max_{\Omega_j}(\boldsymbol{x})$	$\nabla\max_{\Omega_j}(\boldsymbol{x})$
Negative entropy	$\gamma\sum_{i=1}^m y_i \log y_i$	$\gamma\sum_{i=1}^m e^{\frac{x_i}{\gamma}-1}$	$e^{\frac{\boldsymbol{x}}{\gamma}-\mathbf{1}_m}$	$\gamma\log\sum_{i=1}^m e^{\frac{x_i}{\gamma}} - \gamma\log b_j$	$\frac{e^{\frac{\boldsymbol{x}}{\gamma}}}{\sum_{i=1}^m e^{\frac{x_i}{\gamma}}}$
Squared 2-norm	$\frac{\gamma}{2}\|\boldsymbol{y}\|^2$	$\frac{1}{2\gamma}\sum_{i=1}^m [x_i]_+^2$	$\frac{1}{\gamma}[\boldsymbol{x}]_+$	$\boldsymbol{x}^\top\boldsymbol{y}^\star - \frac{\gamma b_j}{2}\|\boldsymbol{y}^\star\|^2$	$\boldsymbol{y}^\star = P_{\triangle^m}\left(\frac{\boldsymbol{x}}{\gamma b_j}\right)$
Group-lasso	$\frac{\gamma}{2}\|\boldsymbol{y}\|^2 + \gamma\mu\sum_{G\in\mathcal{G}}\|\boldsymbol{y}_G\|$	(11)	(12)	No closed form available	

Proposition 2 *Smoothed semi-dual*

$$\mathrm{OT}_\Omega(\boldsymbol{a},\boldsymbol{b}) = \max_{\boldsymbol{\alpha}\in\mathbb{R}^m} \boldsymbol{\alpha}^\top\boldsymbol{a} - \mathrm{OT}^*_\Omega(\boldsymbol{\alpha},\boldsymbol{b}) \qquad (10)$$

The optimal solution $T^\star$ of (5) *can be recovered from $\boldsymbol{\alpha}^\star$ by $\boldsymbol{t}_j^\star = b_j\nabla\max_{\Omega_j}(\boldsymbol{\alpha}^\star - \boldsymbol{c}_j) \quad \forall j\in[n]$.*

Proof. $\mathrm{OT}_\Omega(\boldsymbol{a},\boldsymbol{b})$ is a closed and convex function of $\boldsymbol{a}$. Therefore, $\mathrm{OT}_\Omega(\boldsymbol{a},\boldsymbol{b}) = \mathrm{OT}^{**}_\Omega(\boldsymbol{a},\boldsymbol{b})$. ☐

We can interpret this semi-dual as a variant of (4), where the max operator has been replaced with its smoothed counterpart, $\max_{\Omega_j}$. Note that $\boldsymbol{\alpha}^\star$, as obtained by solving the smoothed dual (7) or semi-dual (10), is the gradient of $\mathrm{OT}_\Omega(\boldsymbol{a},\boldsymbol{b})$ w.r.t. $\boldsymbol{a}$ when $\boldsymbol{\alpha}^\star$ is unique or a sub-gradient otherwise. This is useful when learning with OT_Ω as a loss, as done with entropic regularization in [Frogner et al., 2015].

Solving the optimization problems. The dual and semi-dual we derived are unconstrained, differentiable and concave optimization problems. They can therefore be solved using gradient-based algorithms, as long as we know how to compute $\nabla\delta_\Omega$ and $\nabla\max_\Omega$. In our experiments, we use L-BFGS [Liu and Nocedal, 1989], for both the dual and semi-dual formulations.

3.3 Closed-form expressions

We derive in this section closed-form expressions for δ_Ω, $\max_\Omega$ and their gradients for specific choices of Ω.

Negative entropy. We choose $\Omega(\boldsymbol{y}) = -\gamma H(\boldsymbol{y})$, where $H(\boldsymbol{y}) \coloneqq -\sum_i y_i \log y_i$ is the entropy. For that choice, we get analytical expressions for δ_Ω, $\max_\Omega$ and their gradients (cf. Table 1). Since Ω is γ-strongly convex w.r.t. the 1-norm over $\triangle^m$ [Shalev-Shwartz, 2007, Lemma 16], $\max_\Omega$ is $\frac{1}{\gamma}$-strongly smooth w.r.t. the ∞-norm. However, since Ω is only strictly convex over $\mathbb{R}^m_{>0}$, δ_Ω is differentiable but *not* smooth. The dual and semi-dual with this Ω were derived in [Cuturi and Doucet, 2014] and [Genevay et al., 2016], respectively.

Next, we present two choices of Ω that induce sparsity in transportation plans. The resulting dual and semi-dual expressions are new, to our knowledge.

Squared 2-norm. We choose $\Omega(\boldsymbol{y}) = \frac{\gamma}{2}\|\boldsymbol{y}\|^2$. We again obtain closed-form expressions for δ_Ω, $\max_\Omega$ and their gradients (cf. Table 1). Since Ω is γ-strongly convex w.r.t. the 2-norm over $\mathbb{R}^m$, both δ_Ω and $\max_\Omega$ are $\frac{1}{\gamma}$-strongly smooth w.r.t. the 2-norm. Projecting a vector onto the simplex, as required to compute $\max_\Omega$ and its gradient, can be done exactly in worst-case $O(m\log m)$ time using the algorithm of [Michelot, 1986] and in expected $O(m)$ time using the randomized pivot algorithm of [Duchi et al., 2008]. Squared 2-norm regularization can output exactly sparse transportation plans (the primal-dual relationship for (7) is $t^\star_{i,j} = \frac{1}{\gamma}[\alpha_i^\star + \beta_j^\star - c_{i,j}]_+$) and is numerically stable without any particular implementation trick.

Group lasso. Courty et al. [2016] recently proposed to use $\Omega(\boldsymbol{y}) = \gamma(\sum_i y_i\log y_i + \mu\sum_{G\in\mathcal{G}}\|\boldsymbol{y}_G\|)$, where $\boldsymbol{y}_G$ denotes the subvector of $\boldsymbol{y}$ restricted to the set G, and showed that this regularization improves accuracy in the context of domain adaptation. Since Ω includes a negative entropy term, the same remarks as for negative entropy apply regarding the differentiability of δ_Ω and smoothness of $\max_\Omega$. Unfortunately, a closed-form solution is available for neither (6) nor (8). However, since the log keeps $\boldsymbol{y}$ in the strictly positive orthant and $\|\boldsymbol{y}_G\|$ is differentiable everywhere in that orthant, we can use any proximal gradient algorithm to solve these problems to arbitrary precision.

A drawback of this choice of Ω, however, is that group sparsity is never truly achieved. To address this issue, we propose to use $\Omega(\boldsymbol{y}) = \gamma(\frac{1}{2}\|\boldsymbol{y}\|^2 + \mu\sum_{G\in\mathcal{G}}\|\boldsymbol{y}_G\|)$ instead. For that choice, δ_Ω is smooth and is equal to

$$\delta_\Omega(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{y}^\star - \Omega(\boldsymbol{y}^\star), \qquad (11)$$

where $\boldsymbol{y}^\star$ decomposes over groups $G\in\mathcal{G}$ and equals

$$\boldsymbol{y}_G^\star = \operatorname*{argmin}_{\boldsymbol{y}_G\geq 0} \frac{1}{2}\|\boldsymbol{y}_G - \boldsymbol{x}_G/\gamma\|^2 + \mu\|\boldsymbol{y}_G\| = \nabla\delta_\Omega(\boldsymbol{x})_G. \qquad (12)$$

As noted in the context of group-sparse NMF [Kim et al., 2012], (12) admits a closed-form solution

$$\boldsymbol{y}_G^\star = \operatorname*{argmin}_{\boldsymbol{y}_G} \frac{1}{2}\left\|\boldsymbol{y}_G - \boldsymbol{x}_G^+\right\|^2 + \mu\|\boldsymbol{y}_G\| = \left[1 - \frac{\mu}{\|\boldsymbol{x}_G^+\|}\right]_+ \boldsymbol{x}_G^+,$$

where we defined $\boldsymbol{x}^+ := \frac{1}{\gamma}[\boldsymbol{x}]_+$. We have thus obtained an efficient way to compute exact gradients of δ_Ω, making it possible to solve the dual using gradient-based algorithms. In contrast, Courty et al. [2016] use a generalized conditional gradient algorithm whose iterations require expensive calls to Sinkhorn. Finally, because $\boldsymbol{t}_j^\star = \nabla\delta_\Omega(\boldsymbol{\alpha}^\star + \beta_j^\star \mathbf{1}_m - \boldsymbol{c}_j)\ \forall j \in [n]$, the obtained transportation plan will be truly group-sparse.

4 Relaxed primal $\leftrightarrow$ Strong dual

We now explore the opposite way to define smooth OT problems while retaining sparse transportation plans: replace marginal constraints in the primal with approximate constraints. When relaxing both marginal constraints, we define the next formulation:

Definition 2 *Relaxed smooth primal*

$$\mathrm{ROT}_\Phi(\boldsymbol{a},\boldsymbol{b}) \coloneqq \min_{T\geq 0} \langle T, C\rangle + \frac{1}{2}\Phi(T\mathbf{1}_n, \boldsymbol{a}) + \frac{1}{2}\Phi(T^\top \mathbf{1}_m, \boldsymbol{b}), \tag{13}$$

where $\Phi(\boldsymbol{x},\boldsymbol{y})$ is a smooth divergence measure.

We may also relax only one of the marginal constraints:

Definition 3 *Semi-relaxed smooth primal*

$$\widetilde{\mathrm{ROT}}_\Phi(\boldsymbol{a},\boldsymbol{b}) \coloneqq \min_{\substack{T\geq 0\\ T^\top \mathbf{1}_m = \boldsymbol{b}}} \langle T, C\rangle + \Phi(T\mathbf{1}_n, \boldsymbol{a}) \tag{14}$$

where $\Phi(\boldsymbol{x},\boldsymbol{y})$ is defined as in Definition 2.

For both (13) and (14), the transportation plans will be typically sparse. As discussed in more details in §7, these formulations are similar to [Frogner et al., 2015, Chizat et al., 2016], with the key difference that we do not regularize T with an entropic term. In addition, for Φ, we propose to use $\Phi(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2\gamma}\|\boldsymbol{x}-\boldsymbol{y}\|^2$, which is $\frac{1}{\gamma}$-smooth, while these works use a generalized Kullback-Leibler (KL) divergence, which is not smooth. Relaxing the marginal constraints is useful when normalizing input measures to unit mass is not suitable [Gramfort et al., 2015] or to allow for only partial displacement of mass. Relaxing only one of the two constraints is useful in color transfer [Rabin et al., 2014], where we would like all the probability mass of the source image to be accounted for but not necessarily for the reference image.

Dual interpretation. As we show in Appendix A.3, in the case $\Phi(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2\gamma}\|\boldsymbol{x}-\boldsymbol{y}\|^2$, the dual of (13) can be interpreted as the original dual with additional squared 2-norm regularization on the dual variables $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. For the dual of (14), the additional regularization is on $\boldsymbol{\alpha}$ only (on the original dual or equivalently on the original semi-dual). For that choice of Φ, the duals of (13) and (14) are strongly convex. The dual formulations are crucial to derive our bounds in §5.

Solving the optimization problems. While the relaxed and semi-relaxed primals (13) and (14) are still constrained problems, it is much easier to project on their constraint domain than on $\mathcal{U}(\boldsymbol{a},\boldsymbol{b})$. For the relaxed primal, in our experiments we use L-BFGS-B, a variant of L-BFGS suitable for box-constrained problems [Byrd et al., 1995]. For the semi-relaxed primal, we use FISTA [Beck and Teboulle, 2009]. Since the constraint domain of (14) has the structure of a Cartesian product $b_1\triangle^m \times \cdots \times b_n \triangle^m$, we can easily project any T on it by column-wise projection on the (scaled) simplex. Although not exlored in this paper, the block Frank-Wolfe algorithm [Lacoste-Julien et al., 2012] is also a good fit for the semi-relaxed primal.

5 Theoretical bounds

Convergence rates. The dual (7) is not smooth in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ when using entropic regularization but it is when using the squared 2-norm, with constant upper-bounded by $^n/_\gamma$ w.r.t. $\boldsymbol{\alpha}$ and $^m/_\gamma$ w.r.t. $\boldsymbol{\beta}$. The semi-dual (10) is smooth for both regularizations, with the same constant of $^1/_\gamma$, albeit not in the same norm. The relaxed and semi-relaxed primals (13) and (14) are both $^1/_\gamma$-smooth when using $\Phi(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2\gamma}\|\boldsymbol{x}-\boldsymbol{y}\|^2$. However, none of these problems are strongly convex. From standard convergence analysis of (projected) gradient descent for smooth but non-strongly convex problems, the number of iterations to reach an ϵ-accurate solution w.r.t. the smoothed problems is $O(^1/_{\gamma\epsilon})$ or $O(^1/_{\sqrt{\gamma\epsilon}})$ with Nesterov acceleration.

Approximation error. Because the smoothed problems approach unregularized OT as $\gamma \to 0$, there is a trade-off between convergence rate w.r.t. the smoothed problem and approximation error w.r.t. unregularized OT. A question is then which smoothed formulations and which regularizations have better approximation error. Our first theorem bounds $\mathrm{OT}_\Omega - \mathrm{OT}$ in the case of entropic and squared 2-norm regularization.

Theorem 1 *Approximation error of OT_Ω*

Let $\boldsymbol{a} \in \triangle^m$ and $\boldsymbol{b} \in \triangle^n$. Then,

$$\gamma L \leq \mathrm{OT}_\Omega(\boldsymbol{a},\boldsymbol{b}) - \mathrm{OT}(\boldsymbol{a},\boldsymbol{b}) \leq \gamma U,$$

where we defined L and U as follows.

Ω	Neg. entropy	Squared 2-norm
L	$-H(\boldsymbol{a}) - H(\boldsymbol{b})$	$\frac{1}{2}\sum_{i,j=1}^{m,n}\left(\frac{a_i}{n} + \frac{b_j}{m} - \frac{1}{mn}\right)^2$
U	$-\max\{H(\boldsymbol{a}), H(\boldsymbol{b})\}$	$\frac{1}{2}\min\left\{\|\boldsymbol{a}\|^2, \|\boldsymbol{b}\|^2\right\}$

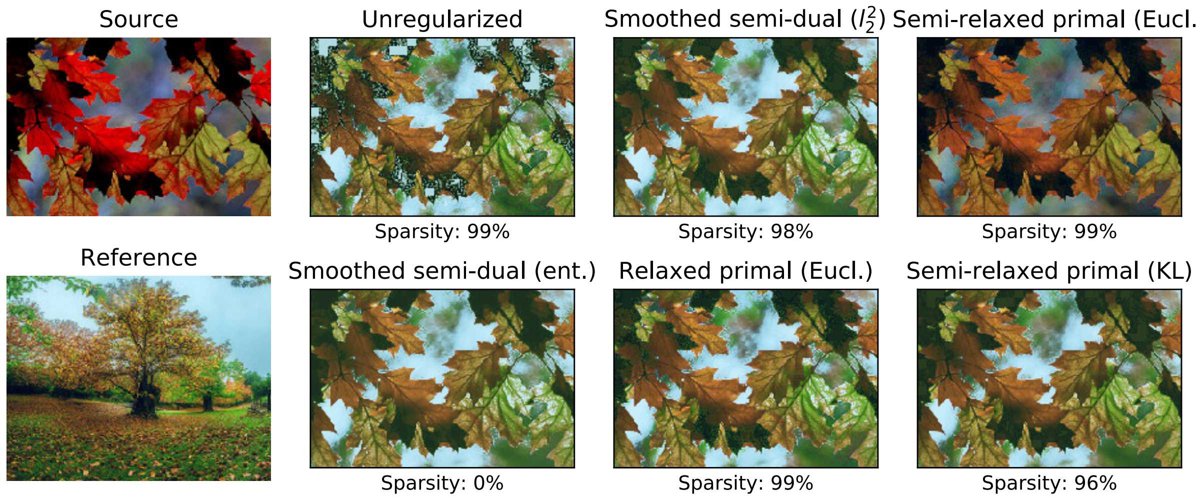

Figure 2: Result comparison for different formulations on the task of color transfer. For regularized formulations, we solve the optimization problem with $\gamma \in \{10^{-4}, 10^{-2}, \ldots, 10^{4}\}$ and choose the most visually pleasing result. The sparsity indicated below each image is the percentage of zero elements in the optimal transportation plan.

Proof is given in Appendix A.4. Our result suggests that, for the same γ, the approximation error can often be smaller with squared 2-norm than with entropic regularization. In particular, this is true whenever $\min\{H(\boldsymbol{a}), H(\boldsymbol{b})\} > \frac{1}{2}\min\left\{\|\boldsymbol{a}\|^2, \|\boldsymbol{b}\|^2\right\}$, which is often the case in practice since $0 \le \min\{H(\boldsymbol{a}), H(\boldsymbol{b})\} \le \min\{\log m, \log n\}$ while $0 \le \frac{1}{2}\min\left\{\|\boldsymbol{a}\|^2, \|\boldsymbol{b}\|^2\right\} \le \frac{1}{2}$. Our second theorem bounds $\mathrm{OT} - \mathrm{ROT}_\Phi$ and $\mathrm{OT} - \widetilde{\mathrm{ROT}}_\Phi$ when Φ is the squared Euclidean distance.

Theorem 2 *Approximation error of ROT_Φ, $\widetilde{\mathrm{ROT}}_\Phi$*
Let $\boldsymbol{a} \in \triangle^m$, $\boldsymbol{b} \in \triangle^n$, $\Phi(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2\gamma}\|\boldsymbol{x} - \boldsymbol{y}\|^2$. *Then,*

$$0 \le \mathrm{OT}(\boldsymbol{a}, \boldsymbol{b}) - \mathrm{ROT}_\Phi(\boldsymbol{a}, \boldsymbol{b}) \le \gamma L$$
$$0 \le \mathrm{OT}(\boldsymbol{a}, \boldsymbol{b}) - \widetilde{\mathrm{ROT}}_\Phi(\boldsymbol{a}, \boldsymbol{b}) \le \gamma \widetilde{L}$$

where we defined

$$L := \|C\|_\infty^2 \min\{\nu_1 + n, \nu_2 + m\}^2$$
$$\nu_1 := \max\left\{(2 + n/m)\left\|\boldsymbol{a}^{-1}\right\|_\infty, \ \left\|\boldsymbol{b}^{-1}\right\|_\infty\right\}$$
$$\nu_2 := \max\left\{\left\|\boldsymbol{a}^{-1}\right\|_\infty, \ (2 + m/n)\left\|\boldsymbol{b}^{-1}\right\|_\infty\right\}$$
$$\widetilde{L} := 2\|C\|_\infty^2 \left\|\boldsymbol{a}^{-1}\right\|_\infty^2.$$

Proof is given in Appendix A.5. While the bound for $\widetilde{\mathrm{ROT}}_\Phi$ is better than that of ROT_Φ, both are worse than that of OT_Ω, suggesting that the smoothed dual formulations are the way to go when low approximation error w.r.t. unregularized OT is important.

6 Experimental results

We showcase our formulations on color transfer, which is a classical OT application [Pitié et al., 2007]. More experimental results are presented in Appendix C.

6.1 Application to color transfer

Experimental setup. Given an image of size $u \times v$, we represent its pixels in RGB color space. We apply k-means clustering to quantize the image down to m colors. This produces m color centroids $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathbb{R}^3$. We can count how many pixels were assigned to each centroid and normalizing by uv gives us a color histogram $\boldsymbol{a} \in \triangle^m$. We repeat the same process with a second image to obtain $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \mathbb{R}^3$ and $\boldsymbol{b} \in \triangle^n$. Next, we apply any of the proposed methods with cost matrix $c_{i,j} = d(\boldsymbol{x}_i, \boldsymbol{y}_j)$, where d is some discrepancy measure, to obtain a (possibly relaxed) transportation plan $T \in \mathbb{R}_+^{m \times n}$. For each color centroid $\boldsymbol{x}_i$, we apply a barycentric projection to obtain a new color centroid

$$\hat{\boldsymbol{x}}_i := \operatorname*{argmin}_{\boldsymbol{x} \in \mathbb{R}^3} \sum_{j=1}^{n} t_{i,j}\, d(\boldsymbol{x}, \boldsymbol{y}_j).$$

When $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^2$, as used in our experiments, the above admits a closed-form solution: $\hat{\boldsymbol{x}}_i = \frac{\sum_{j=1}^n t_{i,j}\boldsymbol{y}_j}{\sum_{j=1}^n t_{i,j}}$. Finally, we use the new color $\hat{\boldsymbol{x}}_i$ for all pixels assigned to $\boldsymbol{x}_i$. The same process can be performed with respect to the $\boldsymbol{y}_j$, in order to transfer the colors in the other direction. We use two public domain images "fall foliage" by Bernard Spragg and "comunion" by Abel Maestro Garcia, and reduce the number of colors to $m = n = 4096$. We compare smoothed dual approaches and (semi-)relaxed primal approaches. For the semi-relaxed primal, we also compared with $\Phi(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{\gamma}\mathrm{KL}(\boldsymbol{x}\|\boldsymbol{y})$, where $\mathrm{KL}(\boldsymbol{x}\|\boldsymbol{y})$ is the generalized KL divergence, $\boldsymbol{x}^\top \log\left(\frac{\boldsymbol{x}}{\boldsymbol{y}}\right) - \boldsymbol{x}^\top \mathbf{1} +$

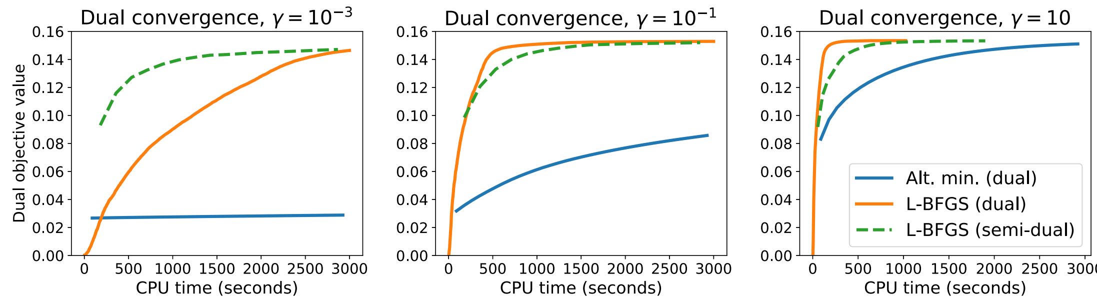

Figure 3: Solver comparison for the smoothed dual and semi-dual, with squared 2-norm regularization. With $\gamma = 10$, which was also the best value selected in Figure 2, the maximum is reached in less than 4 minutes.

$\boldsymbol{y}^\top \mathbf{1}$. This choice is differentiable but not smooth. We ran the aforementioned solvers for up to 1000 epochs.

Results. Our results are presented in Figure 2. All formulations clearly produced better results than unregularized OT. With the exception of the entropy-smoothed semi-dual formulation, all formulations produced extremely sparse transportation plans. The semi-relaxed primal formulation with Φ set to the squared Euclidean distance was the only one to produce colors with a darker tone.

6.2 Solver and objective comparison

We compared the smoothed dual and semi-dual when using squared 2-norm regularization. In addition to L-BFGS on both objectives, we also compared with alternating minimization in the dual. As we show in Appendix B, exact block minimization w.r.t. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be carried out by projection onto the simplex.

Results. We ran the comparison using the same data as in §6.1. Results are indicated in Figure 3. When the problem is loosely regularized, we made two key findings: i) L-BFGS converges much faster in the semi-dual than in the dual, ii) alternating minimization converges extremely slowly. The reason for i) could be the better smoothness constant of the semi-dual (cf. §5). Since alternating minimization and the semi-dual have roughly the same cost per iteration (cf. Appendix B), the reason for ii) is *not* iteration cost but a convergence issue of alternating minimization. When using larger regularization, L-BFGS appears to converge slighly faster on the dual than on the semi-dual, which is likely thanks to its cheap-to-compute gradients.

6.3 Approximation error comparison

We compared empirically the approximation error of smoothed formulations w.r.t. unregularized OT according to four criteria: transportation plan error, marginal constraint error, value error and regularized value error (cf. Figure 4 for a precise definition). For the dual approaches, we solved the smoothed semi-dual objective (10), since, as we discussed in §5, it has the same smoothness constant of $1/\gamma$ for both entropic and squared 2-norm regularizations, implying similar convergence rates in theory. In addition, in the case of entropic regularization, the expressions of $\max_\Omega$ and $\nabla \max_\Omega$ are trivial to stabilize numerically using standard log-sum-exp implementation tricks.

Results. We ran the comparison using the same data as in §6.1. Results are indicated in Figure 4. For the transportation plan error and the (regularized) value error, entropic regularization required 100 times smaller γ to achieve the same error. This confirms, as suggested by Theorem 1, that squared 2-norm regularization is typically tighter. Unsurprisingly, the semi-relaxed primal was tighter than the relaxed primal in all four criteria. A runtime comparison of smoothed formulations is also important. However, a rigorous comparison would require carefully engineered implementations and is therefore left for future work.

7 Related work

Regularized OT. Problems similar to (5) for general Ω were considered in [Dessein et al., 2016]. Their work focuses on strictly convex and differentiable Ω for which there exists an associated Bregman divergence. Following [Benamou et al., 2015], they show that (5) can then be reformulated as a Bregman projection onto the transportation polytope and solved using Dykstra's algorithm [1985]. While Dykstra's algorithm can be interpreted implicitly as a two-block alternating minimization scheme on the dual problem, neither the dual nor the semi-dual expressions were derived. These expressions allow us to make use of arbitrary solvers, including quasi-Newton ones like L-BFGS, which as we showed empirically, converge much faster on loosely regularized problems. Our framework can also accomodate non-differentiable regularizations for which there does not exist an associated Bregman divergence, such as those that include a group lasso

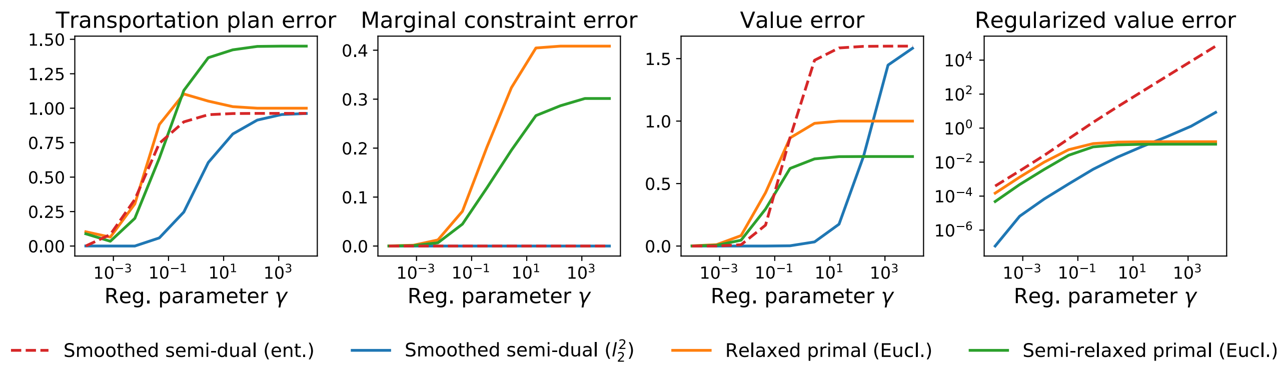

Figure 4: Approximation error w.r.t. unregularized OT empirically achieved by different smoothed formulations on the task of color transfer. Let $T^\star$ be en optimal solution of the unregularized LP (2) and $T^\star_\gamma$ be an optimal solution of one of the smoothed formulations with regularization parameter γ. The transportation plan error is $\|T^\star_\gamma - T^\star\|/\|T^\star\|$. The marginal constraint error is $\|T^\star_\gamma \mathbf{1}_n - \boldsymbol{a}\| + \|(T^\star_\gamma)^\top \mathbf{1}_m - \boldsymbol{b}\|$. The value error is $|\langle T^\star_\gamma, C\rangle - \langle T^\star, C\rangle|/\langle T^\star, C\rangle$. The regularized value error is $|v - \langle T^\star, C\rangle|$, where v is one of $\mathrm{OT}_\Omega(\boldsymbol{a},\boldsymbol{b})$, $\mathrm{ROT}_\Phi(\boldsymbol{a},\boldsymbol{b})$ and $\widetilde{\mathrm{ROT}}_\Phi(\boldsymbol{a},\boldsymbol{b})$. For the regularized value error, our empirical findings confirm what Theorem 1 suggested, namely that, for the same value of γ, squared 2-norm regularization is quite tighter than entropic regularization.

term. Squared 2-norm regularization was recently considered in [Li et al., 2016] as well as in [Essid and Solomon, 2017] but for a reformulation of the Wasserstein distance of order 1 as a min cost flow problem along the edges of a graph.

Relaxed OT. There has been a large number of proposals to extend OT to unbalanced positive measures. Static formulations with approximate marginal constraints based on the KL divergence have been proposed in [Frogner et al., 2015, Chizat et al., 2016]. The main difference with our work is that these formulations include an additional entropic regularization on T. While this entropic term enables a Sinkhorn-like algorithm, it also prevents from obtaining sparse T and requires the tuning of an additional hyper-parameter. Relaxing only one of the two marginal constraints with an inequality was investigated for color transfer in [Rabin et al., 2014]. Benamou [2003] considered an interpolation between OT and squared Euclidean distances:

$$\min_{\boldsymbol{x} \in \triangle^m} \mathrm{OT}(\boldsymbol{x}, \boldsymbol{b}) + \frac{1}{2\gamma}\|\boldsymbol{x} - \boldsymbol{a}\|^2. \tag{15}$$

While on first sight this looks quite different, this is in fact equivalent to our semi-relaxed primal formulation when $\Phi(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2\gamma}\|\boldsymbol{x}-\boldsymbol{y}\|^2$ since (15) is equal to

$$\begin{aligned} &\min_{\boldsymbol{x} \in \triangle^m} \min_{\substack{T \geq 0 \\ T\mathbf{1}_n = \boldsymbol{x} \\ T^\top \mathbf{1}_m = \boldsymbol{b}}} \langle T, C\rangle + \frac{1}{2\gamma}\|\boldsymbol{x} - \boldsymbol{a}\|^2 \\ = &\min_{\substack{T \geq 0 \\ T^\top \mathbf{1}_m = \boldsymbol{b}}} \langle T, C\rangle + \frac{1}{2\gamma}\|T\mathbf{1}_n - \boldsymbol{a}\|^2 = \widetilde{\mathrm{ROT}}_\Phi(\boldsymbol{a},\boldsymbol{b}). \end{aligned}$$

However, the bounds in §5 are to our knowledge new. A similar formulation but with a group-lasso penalty on T instead of $\frac{1}{2\gamma}\|T\mathbf{1}_n - \boldsymbol{a}\|^2$ was considered in the context of convex clustering [Carli et al., 2013].

Smoothed LPs. Smoothed linear programs have been investigated in other contexts. The two closest works to ours are [Meshi et al., 2015b] and [Meshi et al., 2015a], in which smoothed LP relaxations based on the squared 2-norm are proposed for maximum a-posteriori inference. One innovation we make compared to these works is to abstract away the regularization by introducing the δ_Ω and $\max_\Omega$ functions.

8 Conclusion

We proposed in this paper to regularize both the primal and dual OT formulations with a strongly convex term, and showed that this corresponds to relaxing the dual and primal constraints with smooth approximations. There are several important avenues for future work. The conjugate expression (9) should be useful for barycenter computation [Cuturi and Peyré, 2016] or dictionary learning [Rolet et al., 2016] with squared 2-norm instead of entropic regularization. On the theoretical side, while we provided convergence guarantees w.r.t. the OT distance *value* as the regularization vanishes, which suggested the advantage of squared 2-norm regularization, it would also be important to study the convergence w.r.t. the *transportation plan*, as was done for entropic regularization by Cominetti and San Martín [1994]. Finally, studying optimization algorithms that can cope with large-scale data is important. We believe SAGA [Defazio et al., 2014] is a good candidate since it is stochastic, supports proximity operators, is adaptive to non-strongly convex problems and can be parallelized [Leblond et al., 2017].

Acknowledgements

We thank Arthur Mensch and the anonymous reviewers for constructive comments.

References

On the transfer of masses (in russian). *Doklady Akademii Nauk*, 37(2):227–229, 1942.

Jason Altschuler, Jonathan Wee, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Proc. of NIPS*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proc. of ICML*, volume 70, pages 214–223, 2017.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2 (1):183–202, 2009.

Jean-David Benamou. Numerical resolution of an unbalanced mass transport problem. *ESAIM: Mathematical Modelling and Numerical Analysis*, 37(5): 851–868, 2003.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2): A1111–A1138, 2015.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

Gilberto Calvillo and David Romero. On the closest point to the origin in transportation polytopes. *Discrete Appl. Math.*, 210:88–102, 2016.

Francesca P Carli, Lipeng Ning, and Tryphon T Georgiou. Convex clustering via optimal mass transport. *arXiv preprint arXiv:1307.5459*, 2013.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced transport problems. *arXiv preprint arXiv:1607.05816*, 2016.

Roberto Cominetti and Jaime San Martín. Asymptotic analysis of the exponential penalty trajectory in linear programming. *Mathematical Programming*, 67(1-3):169–187, 1994.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 2016.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2006.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proc. of NIPS*, pages 2292–2300. 2013.

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693, 2014.

Marco Cuturi and Gabriel Peyré. A smoothed dual approach for variational wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

John M Danskin. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14 (4):641–664, 1966.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Proc. of NIPS*, pages 1646–1654, 2014.

Arnaud Dessein, Nicolas Papadakis, and Jean-Luc Rouas. Regularized optimal transport and the rot mover's distance. *arXiv preprint arXiv:1610.06447*, 2016.

John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *Proc. of ICML*, 2008.

Richard L Dykstra. An iterative procedure for obtaining i-projections onto the intersection of convex sets. *The annals of Probability*, pages 975–984, 1985.

Montacer Essid and Justin Solomon. Quadratically-regularized optimal transport on graphs. *arXiv preprint arXiv:1704.08200*, 2017.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. In *Proc. of NIPS*, pages 2053–2061, 2015.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Proc. of NIPS*, pages 3440–3448. 2016.

Alexandre Gramfort, Gabriel Peyré, and Marco Cuturi. Fast optimal transport averaging of neuroimaging data. In *Proc. of International Conference on Information Processing in Medical Imaging*, pages 261–272, 2015.

Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *Journal of Machine Learning Research*, 13:1865–1890, 2012.

Jingu Kim, Renato DC Monteiro, and Haesun Park. Group sparsity in nonnegative matrix factorization. In *Proc. of SIAM International Conference on Data Mining*, pages 851–862, 2012.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *Proc. of ICML*, pages 957–966, 2015.

Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *Proc. of ICML*, 2012.

Rmi Leblond, Fabian Pedregosa, and Simon Lacoste-Julien. ASAGA: Asynchronous Parallel SAGA. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 46–54, 2017.

Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete and Continuous Dynamical Systems*, 34:1879–1920, 2012.

Wuchen Li, Stanley Osher, and Wilfrid Gangbo. A fast algorithm for earth mover's distance based on optimal transport and l1 type regularization. *arXiv preprint arXiv:1609.07092*, 2016.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

Ofer Meshi, Amir Globerson, and Tommi S. Jaakkola. Convergence rate analysis of map coordinate minimization algorithms. In *Proc. of NIPS*, pages 3014–3022. 2012.

Ofer Meshi, Mehrdad Mahdavi, and Alexander G. Schwing. Smooth and strong: MAP inference with linear convergence. In *Proc. of NIPS*, 2015a.

Ofer Meshi, Nathan Srebro, and Tamir Hazan. Efficient training of structured svms via soft constraints. In *Proc. of AISTATS*, pages 699–707, 2015b.

Christian Michelot. A finite algorithm for finding the projection of a point onto the canonical simplex of $\mathbb{R}^n$. *Journal of Optimization Theory and Applications*, 50(1):195–200, 1986.

Boris Muzellec, Richard Nock, Giorgio Patrini, and Frank Nielsen. Tsallis regularized optimal transport and ecological inference. In *AAAI*, pages 2387–2393, 2017.

Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport*. 2017.

François Pitié, Anil C Kokaram, and Rozenn Dahyot. Automated colour grading using colour distribution transfer. *Computer Vision and Image Understanding*, 107(1):123–137, 2007.

Julien Rabin, Sira Ferradans, and Nicolas Papadakis. Adaptive color transfer with relaxed optimal transport. In *Proc. of International Conference on Image Processing*, pages 4852–4856, 2014.

Antoine Rolet, Marco Cuturi, and Gabriel Peyré. Fast dictionary learning with a smoothed wasserstein loss. In *Proc. of AISTATS*, pages 630–638, 2016.

David Romero. Easy transportation-like problems on k-dimensional arrays. *Journal of Optimization Theory and Applications*, 66:137–147, 1990.

Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *arXiv preprint arXiv:1610.06519*, 2016.

Erwin Schrödinger. Über die umkehrung der naturgesetze. *Phys. Math.*, 144:144–153, 1931.

Shai Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University of Jerusalem, 2007.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Justin Solomon, Raif Rustamov, Guibas Leonidas, and Adrian Butscher. Wasserstein propagation for semi-supervised learning. In *Proc. of ICML*, pages 306–314, 2014.

Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.