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
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., $t$, $\mathbf{t}$ and $T$, respectively. Given a matrix $T$, we denote its elements by $t_{ij}$ and its columns by $\mathbf{t}_j$. 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 $\Delta^m := \{ \mathbf{y} \in \mathbb{R}^m : \| \mathbf{y} \|_1 = 1 \}$ and the Euclidean projection onto it by $P_{\Delta^m}(\mathbf{x}) := \arg\min_{\mathbf{y} \in \Delta^m} \| \mathbf{y} - \mathbf{x} \|_2^2$. We denote $[\mathbf{x}]_+ := \max(\mathbf{x}, \mathbf{0})$, 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

$$f^*(\mathbf{x}) := \sup_{\mathbf{y} \in \text{dom } f} \mathbf{y}^\top \mathbf{x} - f(\mathbf{y}).$$

(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 $f^*$:

$$\nabla f^*(\mathbf{x}) = \arg\max_{\mathbf{y} \in \text{dom } f} \mathbf{y}^\top \mathbf{x} - f(\mathbf{y}).$$

The dual of a norm $\| \cdot \|$ is defined by $\| \cdot \|_* := \sup_{\| \mathbf{y} \| \leq 1} \mathbf{y}^\top \mathbf{x}$. We say that a function is $\gamma$-smooth w.r.t. a norm $\| \cdot \|$ if it is differentiable everywhere and its gradient is $\gamma$-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 $\gamma$-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 $\mathbf{a} \in \Delta^m$ and $\mathbf{b} \in \Delta^n$. Rather than performing a point-wise 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

![Image of transportation plans comparison](https://github.com/mblondel/smooth-ot)
Kantorovich [1942], is cast as a linear program (LP):
\[
\text{OT}(a, b) := \min_{T \in U(a, b)} \langle T, C \rangle,
\]
where \( U(a, b) \) is the transportation polytope
\[
U(a, b) := \{ T \in \mathbb{R}_+^{m \times n} : T1_n = a, T^\top 1_m = b \}
\]
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 \( a \) and \( b \). Without loss of generality, we will assume \( a > 0 \) and \( b > 0 \) throughout this paper (if \( a_i = 0 \) or \( b_j = 0 \) then the \( i \)-th row or the \( j \)-th column of \( T^* \) is zero). When \( n = m \) and \( C \) is a distance matrix raised to the power \( p \), \( \text{OT}(\cdot, \cdot)^{\frac{1}{p}} \) is a distance on \( \Delta^n \), called the \( \text{Kantorovich} \) [1942], is cast as a linear program (LP):
\[
\text{OT}(a, b) := \max_{\alpha, \beta \in P(C)} \alpha^\top a + \beta^\top b,
\]
where \( P(C) := \{ \alpha \in \mathbb{R}^m, \beta \in \mathbb{R}^n : \alpha_i + \beta_j \leq c_{i,j} \} \). Keeping \( \alpha \) fixed, an optimal solution w.r.t. \( \beta \) 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”
\[
\text{OT}(a, b) = \max_{\alpha \in \mathbb{R}^m} \alpha^\top a - \sum_{j=1}^n b_j \max_{i \in [m]} (\alpha_i - c_{i,j}).
\]
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
\[
\text{OT}_\Omega(a, b) := \min_{T \in U(a, b)} \langle T, C \rangle + \sum_{j=1}^n \Omega(t_j),
\]
where we assume that \( \Omega \) is strongly convex over the intersection of \( \text{dom} \Omega \) and either \( \mathbb{R}_+^m \) or \( \Delta^m \).

These assumptions are sufficient for [3] to be strongly convex w.r.t. \( T \in U(a, b) \). On first sight, solving [3] 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
\[
\delta(x) := \begin{cases} 0, & \text{if } x \leq 0 \\ \infty, & \text{o.w.} \end{cases} = \sup_{y \geq 0} y^\top x.
\]

To define a smoothed version of \( \delta \), we take the convex conjugate of \( \Omega \), restricted to the non-negative orthant:
\[
\delta_\Omega(x) := \sup_{y \geq 0} y^\top x - \Omega(y).
\]

If \( \Omega \) is \( \gamma \)-strongly convex over \( \mathbb{R}_+^m \cap \text{dom} \Omega \), then \( \delta_\Omega \) is \( \frac{1}{\gamma} \)-smooth and its gradient is \( \nabla \delta_\Omega(x) = y^* \), where \( y^* \) is the supremum of [9]. We next show that \( \delta_\Omega \) plays a crucial role in expressing the dual of [3], which is a smooth optimization problem in \( \alpha \) and \( \beta \).

**Proposition 1 Smooth relaxed dual**
\[
\text{OT}_\Omega(a, b) = \max_{\alpha \in \mathbb{R}^m} \alpha^\top a + \beta^\top b - \sum_{j=1}^n \delta_\Omega(\alpha + \beta_j 1_m - c_j).
\]
The optimal solution \( T^* \) of [3] can be recovered from \( (\alpha^*, \beta^*) \) by \( T^*_j = \nabla \delta_\Omega(\alpha^* + \beta^*_j 1_m - c_j) \) \( \forall j \in [n] \).

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 \( \sum_{j=1}^n \delta_\Omega(\alpha + \beta_j 1_m - c_j) \), are now relaxed with soft ones by substituting \( \delta \) with \( \delta_\Omega \).

3.2 Smoothed semi-dual formulation

We now derive the semi-dual of [3], i.e., the dual [7] with one of the two variables eliminated. Without loss of generality, we proceed to eliminate \( \beta \). To do so, we use the notion of smoothed max operator. Notice that
\[
\max(x) := \max_{i \in [m]} x_i = \sup_{y \in \Delta^m} y^\top x \quad \forall x \in \mathbb{R}^m.
\]

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 \( \Omega \), this time restricted to the simplex:
\[
\max_\Omega(x) := \sup_{y \in \Delta^m} y^\top x - \Omega(y).
\]

If \( \Omega \) is \( \gamma \)-strongly convex over \( \Delta^m \cap \text{dom} \Omega \), then \( \max_\Omega \) is \( \frac{1}{\gamma} \)-smooth and its gradient is defined by \( \nabla \max_\Omega(x) = y^* \), where \( y^* \) is the supremum of [8]. We next show that \( \max_\Omega \) plays a crucial role in expressing the conjugate of \( \text{OT}_\Omega \).

**Lemma 1 Conjugate of \( \text{OT}_\Omega \) w.r.t. its first argument**
\[
\text{OT}_{\Omega}^*(\alpha, b) = \sup_{a \in \Delta^m} \alpha^\top a - \text{OT}_{\Omega}(a, b) = \sum_{j=1}^n b_j \max_{c_j} \Omega_j(y) := \frac{1}{b_j} \Omega_j(b_j y),
\]
where \( \Omega_j(y) := \frac{1}{b_j} \Omega(b_j y) \).

A proof is given in Appendix A.2. With the conjugate, we can now easily express the semi-dual of [3], which involves a smooth optimization problem in \( \alpha \).
Table 1: Closed forms for $\delta_\Omega$ (used in smoothed dual), $\max_{\Omega}$ (used in smoothed semi-dual) and their gradients.

<table>
<thead>
<tr>
<th>$\Omega(y)$</th>
<th>$\delta_\Omega(x)$</th>
<th>$\nabla \delta_\Omega(x)$</th>
<th>$\max_{\Omega}(x)$</th>
<th>$\nabla \max_{\Omega}(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative entropy</td>
<td>$\gamma \sum_{i=1}^{m} y_i \log y_i$</td>
<td>$\gamma \sum_{i=1}^{m} e^{\gamma y_i} - 1$</td>
<td>$\gamma \log \sum_{i=1}^{m} e^{\gamma y_i} - \gamma \log b_j$</td>
<td>$-\frac{e^{\gamma y} - e^{\gamma - 1}}{\sum_{i=1}^{m} e^{\gamma y_i}}$</td>
</tr>
<tr>
<td>Squared 2-norm</td>
<td>$\frac{1}{2} |y|^2$</td>
<td>$\frac{1}{2} \sum_{j=1}^{n} [c_j - \gamma y_j]^2$</td>
<td>$\gamma \sum_{j=1}^{n} [c_j - \gamma y_j]^2$</td>
<td>$\gamma \sum_{j=1}^{n} [c_j - \gamma y_j]^2$</td>
</tr>
<tr>
<td>Group-lasso</td>
<td>$\frac{1}{2} |y|^2 + \gamma \mu \sum_{G \in \mathcal{G}} |y_G|$</td>
<td>$\underline{11}$</td>
<td>$\underline{12}$</td>
<td>No closed form available</td>
</tr>
</tbody>
</table>

**Proposition 2** Smoothed semi-dual

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

The optimal solution $T^*$ of (5) can be recovered from $\alpha^*$ by $t_j^* = b_j \nabla \max_{\Omega}(\alpha^* - c_j)$, $\forall j \in \mathbb{N}$.

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

**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 [Lin 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 $\delta_\Omega$, $\max_{\Omega}$ and their gradients for specific choices of $\Omega$.

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

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

**Squared 2-norm.** We choose $\Omega(y) = \frac{1}{2} \|y\|^2$. We again obtain closed-form expressions for $\delta_\Omega$, $\max_{\Omega}$ and their gradients (cf. Table 1). Since $\Omega$ is $\gamma$-strongly convex w.r.t. the 2-norm over $\mathbb{R}^m$, both $\delta_\Omega$ 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 is smooth and is equal to $\Omega_\gamma = \frac{1}{2}(\alpha_i^* + \beta_j^* - c_{i,j})^+$) and is numerically stable without any particular implementation trick.

**Group lasso.** [Courty et al. 2016] recently proposed to use $\Omega(y) = \gamma(\sum_{i} y_i \log y_i + \mu \sum_{G \in \mathcal{G}} \|y_G\|)$, where $y_G$ denotes the subvector of $y$ restricted to the set $G$, and showed that this regularization improves accuracy in the context of domain adaptation. Since $\Omega$ includes a negative entropy term, the same remarks as for negative entropy apply regarding the differentiability of $\delta_\Omega$ and smoothness of $\max_{\Omega}$. Unfortunately, a closed-form solution is available for neither (6) nor (8).

However, since the log keeps $y$ in the strictly positive orthant and $\|y_G\|$ 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 $\Omega$, however, is that group sparsity is never truly achieved. To address this issue, we propose to use $\Omega(y) = \gamma(\frac{1}{2} \|y\|^2 + \mu \sum_{G \in \mathcal{G}} \|y_G\|)$ instead. For that choice, $\delta_\Omega$ is smooth and is equal to

$$\delta_\Omega(x) = x^\top y^* - \Omega(y^*),$$

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

$$y_G^* = \arg\min_{y_G \geq 0} \frac{1}{2} \|y_G - x_G^*\|^2 + \mu \|y_G\| = \nabla \delta_\Omega(x)_G.$$  (12)

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

$$y_G^* = \arg\min_{y_G} \frac{1}{2} \|y_G - x_G^*\|^2 + \mu \|y_G\| = \left[\frac{1 - \mu}{\|x_G^*\|}\right] x_G^*.$$  (13)
where we defined $x^+ := \frac{1}{2}[x]_+$. We have thus obtained an efficient way to compute exact gradients of $\delta_\Psi$, 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 $t_j^* = \nabla \delta_\Psi(\alpha^* + \beta_j T_{1_m} - e_j)$, $\forall j \in [m]$, the obtained transportation plan will be truly group-sparse.

4 Relaxed primal ↔ 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**

\[
\text{ROT}_\Phi(\alpha, \beta) := \min_{T \geq 0} \langle T, C \rangle + \frac{1}{2} \Phi(T_{1_n}, \alpha)^T \frac{1}{2} \Phi(T^T_{1_m}, \beta),
\]

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

We may also relax only one of the marginal constraints:

**Definition 3 Semi-relaxed smooth primal**

\[
\tilde{\text{ROT}}_\Phi(\alpha, \beta) := \min_{T \geq 0} \langle T, C \rangle + \Phi(T_{1_n}, \alpha)
\]

where $\Phi(x, y)$ 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] and [Chizat et al., 2016], with the key difference that we do not regularize $T$ with an entropic term. In addition, for $\Phi$, we propose to use $\Phi(x, y) = \frac{1}{2\gamma} \|x - y\|^2$, which is $\frac{1}{2}$-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 $\Phi(x, y) = \frac{1}{2\gamma} \|x - y\|^2$, the dual of (13) can be interpreted as the original dual with additional squared 2-norm regularization on the dual variables $\alpha$ and $\beta$. For the dual of (14), the additional regularization is on $\alpha$ only (on the original dual or equivalently on the original semi-dual). For that choice of $\Phi$, 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 $\mathcal{U}(\alpha, \beta)$. 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 $b_1\Delta^n \times \cdots \times b_n\Delta^m$, we can easily project any $T$ on it by column-wise projection on the (scaled) simplex. Although not explored 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 $\alpha$ and $\beta$ when using entropic regularization but it is when using the squared 2-norm, with constant upper-bounded by $n/\gamma$ w.r.t. $\alpha$ and $m/\gamma$ w.r.t. $\beta$. 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 $\Phi(x, y) = \frac{1}{2\gamma} \|x - y\|^2$. 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 $\epsilon$-accurate solution w.r.t. the smoothed problems is $O(1/\epsilon)$ or $O(1/\sqrt{\epsilon})$ 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 $\text{OT}_\Omega - \text{OT}$ in the case of entropic and squared 2-norm regularization.

**Theorem 1 Approximation error of $\text{OT}_\Omega$**

Let $\alpha \in \Delta^m$ and $\beta \in \Delta^n$. Then, $\gamma L \leq \text{OT}_\Omega(\alpha, \beta) - \text{OT}(\alpha, \beta) \leq \gamma U$, where we defined $L$ and $U$ as follows.

\[
\begin{array}{ccc}
\Omega & \text{Neg. entropy} & \text{Squared 2-norm} \\
L & \mathcal{O}(H(\alpha) - H(\beta)) & \frac{1}{2} \sum_{1 \leq i, j \leq m} (a_{ij} + b_{ij} - \frac{1}{mn})^2 \\
U & \max \{H(\alpha), H(\beta)\} & \frac{1}{2} \min \{\|a\|^2, \|b\|^2\}
\end{array}
\]
We showcase our formulations on color transfer, which is a classical OT application [Pitie et al., 2007]. More experimental results are presented in Appendix B.

6 Experimental results

We apply any of the proposed methods with cost matrix $c_{i,j} = d(x_i, y_j)$, where $d$ is some discrepancy measure, to obtain a (possibly relaxed) transportation plan $T \in \mathbb{R}_+^{m \times n}$. For each color centroid $x_i$, we apply a barycentric projection to obtain a new color centroid

$$\hat{x}_i := \underset{x \in \mathbb{R}^3}{\text{argmin}} \sum_{j=1}^n t_{i,j} d(x, y_j).$$

When $d(x, y) = \|x - y\|_2$, as in our experiments, the above admits a closed-form solution: $\hat{x}_i = \frac{\sum_{j=1}^n t_{i,j} y_{i,j}}{\sum_{j=1}^n t_{i,j}}$. Finally, we use the new color $\hat{x}_i$ for all pixels assigned to $x_i$. The same process can be performed with respect to the $y_j$, in order to transfer the colors in the other direction. We use two public domain images “fall foliage” by Bernard Spragg and “communion” 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 $\Phi(x, y) = \frac{1}{\gamma} KL(x||y)$, where $KL(x||y)$ is the generalized KL divergence, $x^\top \log \left( \frac{x}{y} \right) - x^\top 1 + \frac{1}{\gamma} x^\top y$.

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 $\gamma$, the approximation error can often be smaller with squared 2-norm than with entropic regularization. In particular, this is true whenever $\min\{H(a), H(b)\} > \frac{1}{2} \min\{\|a\|^2,\|b\|^2\}$, which is often the case in practice since $0 \leq \min\{H(a), H(b)\} \leq \min\{\log m, \log n\}$ while $0 \leq \frac{1}{2} \min\{\|a\|^2,\|b\|^2\} \leq \frac{1}{2}$. Our second theorem bounds $\text{ROT}$ and $\text{RO} \hat{\text{T}}$ when $\Phi$ is the squared Euclidean distance.

**Theorem 2** Approximation error of $\text{ROT}_\Phi$, $\text{RO} \hat{\text{T}}_\Phi$

Let $a \in \Delta^m$, $b \in \Delta^n$, $\Phi(x, y) = \frac{1}{2\gamma} \|x - y\|^2$. Then,

$$0 \leq \text{ROT}(a, b) - \text{RO} \hat{T}_\Phi(a, b) \leq \gamma L$$

$$0 \leq \text{ROT}(a, b) - \text{RO} \hat{T}_\Phi(a, b) \leq \gamma \tilde{L}$$

where we defined

$$L := \|C\|_\infty^2 \min\{\nu_1 + n, \nu_2 + m\}^2$$

$$\nu_1 := \max\{(2 + n/m) \|a^{-1}\|_\infty, \|b^{-1}\|_\infty\}$$

$$\nu_2 := \max\{\|a^{-1}\|_\infty, (2 + m/n) \|b^{-1}\|_\infty\}$$

$$\tilde{L} := 2 \|C\|_\infty^2 \|a^{-1}\|_\infty^2.$$

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

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. $\alpha$ and $\beta$ 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 slightly 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 $\gamma$ 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 $\Omega$ were considered in Dessein et al., 2016. Their work focuses on strictly convex and differentiable $\Omega$ 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 accommodate non-differentiable regularizations for which there does not exist an associated Bregman divergence, such as those that include a group lasso.
Smooth and Sparse Optimal Transport

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

$$\min_{x \in \Delta^m} \text{OT}(x, b) + \frac{1}{2\gamma} \|x - a\|^2. \quad (15)$$

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

$$\min_{x \in \Delta^m} \min_{T \succeq 0} \langle T, C \rangle + \frac{1}{2\gamma} \|x - a\|^2$$

$$= \min_{\gamma \geq 0} \min_{T \succeq 0} \langle T, C \rangle + \frac{1}{2\gamma} \|T 1_n - a\|^2 = \text{ROT}_\Phi(a, b).$$

However, the bounds in (15) are to our knowledge new. A similar formulation but with a group-lasso penalty on $T$ instead of $\frac{1}{2\gamma} \|T 1_n - a\|^2$ 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 $\delta_\Omega$ 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 Martin, 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


Jingu Kim, Renato DC Monteiro, and Haesun Park. 
Group sparsity in nonnegative matrix factorization. 

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. 


