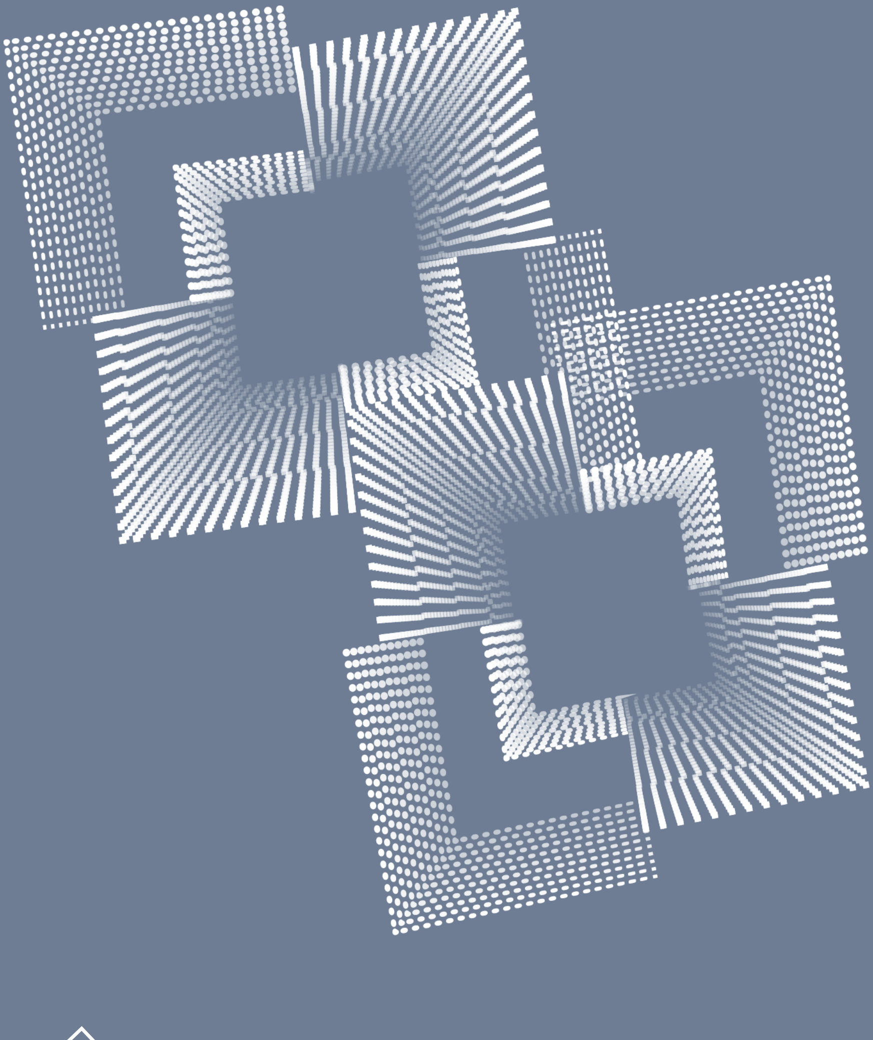
Les Cahiers de la Chaire / N°01

A Smoothed Dual Approach for Variational Wasserstein Problems

Marco Cuturi, Gabriel Peyré



CHAIRE HAVAS-DAUPHINE
ÉCONOMIE ET GESTION DES NOUVELLES DONNÉES

CONNECTING THE DOTS

A SMOOTHED DUAL APPROACH FOR VARIATIONAL WASSERSTEIN PROBLEMS

MARCO CUTURI AND GABRIEL PEYRÉ*

Abstract. Variational problems that involve Wasserstein distances have been recently proposed to summarize and learn from probability measures. Despite being conceptually simple, such problems are computationally challenging because they involve minimizing over quantities (Wasserstein distances) that are themselves hard to compute. We show that the dual formulation of Wasserstein variational problems introduced recently by [13] can be regularized using an entropic smoothing, which leads to smooth, differentiable, convex optimization problems that are simpler to implement and numerically more stable. We illustrate the versatility of this approach by applying it to the computation of Wasserstein barycenters and gradient flows of spacial regularization functionals.

1. Introduction.

To compare two histograms in the probability simplex, information divergences—the Hellinger and χ_2 distances, the Kullback-Leibler and Jensen-Shannon divergences—have the advantages of being simple and fast to compute. Optimal transport distances [34, §7]—*a.k.a.* Wasserstein or earth mover's distances [29]—require more computational effort but are more versatile: by incorporating in their definition a metric between the bins of these histograms, they can compare sparse histograms even if their support do not overlap significantly, which can be crucial when their dimension is large. Their versatility comes, however, at a price: computing optimal transport distances requires solving a costly network flow problem, whose cost scales super-cubicly with the dimension of the considered histograms. That cost becomes even more of a drawback if one attempts to study a family of histograms using the optimal transport geometry.

Despite this computational complexity, optimal transport is becoming increasingly popular in imaging sciences and related fields, such as for instance image retrieval [29, 26], image interpolation [10], computational geometry [23, 21], color image processing [9, 35], image registration [22] and machine learning [15].

1.1. Variational Wasserstein Problems.

Many learning tasks on histograms, such as averaging or clustering them, can be framed as variational problems that involve distances between pairs of histograms. These problems are easily solved when such divergences are Bregman divergences [19, 4, 24], but they are far more challenging when considering instead Wasserstein distances. [2] studied the first problem of this type, the Wasserstein barycenter problem (WBP), and showed that it is related to the multi-marginal optimal transport problem. More recently, [32] proposed the Wasserstein propagation-on-graphs framework and showed that it involves a very large linear program, which can only be feasibly solved for small dimensions and families of histograms. Variational problems that involve Wasserstein distances have, however, the potential to impact a very wide range of applications. Beyond their applicability to unsupervised learning problems and their ramifications into clustering mentioned in [16], they have found usage in statistics to develop population estimators [8], computer graphics to perform image modification [35, 9, 31] and computer vision [36] to summarize complex visual signals.

Beside the computation of barycenters, it is also possible to integrate Wasserstein distances into more general variational problems. For instance, optimal transport

*CNRS and CEREMADE, Université Paris-Dauphine, Place du Maréchal De Lattre De Tassigny, 75775 PARIS CEDEX 16, FRANCE.

distances are used as a data fidelity to perform image denoising [12, 20], image segmentation [28, 33, 30], and Radon transform reconstruction [1, 7].

Our aim in this paper is to propose a computational framework that is both scalable and flexible enough to minimize energies that involve not only Wasserstein distances, but also more general functions such as regularization terms. To do so, we exploit regularization, Legendre duality and the usual toolbox of convex optimization.

1.2. Previous Works. [16] proposes to leverage the entropic regularization of Wasserstein distances introduced by [15] to study the WBP. Their formulation requires, however, to run a numerical subroutine, the Sinkhorn fixed-point iteration, to evaluate these objectives and compute their gradients. On the other hand, [13] show that the Fenchel-Legendre dual of the Wasserstein distance as well as its subgradients can be obtained in *closed form* using nearest-neighbor assignments, that is without having to solve a single optimal transport problem. The authors do, however, struggle with non-differentiable objective functions and use a L-BFGS first order scheme. More recently, [7] have proposed an a generalized version of Sinkhorn's algorithm to compute barycenters based on Bregman's projections. This approach is useful for the barycenter problem, but cannot be easily adapted to solve more advanced variational problems.

A typical use of such more involved variational problems is the approximation of gradient flows. As initially shown by [18], it is indeed possible to approximate solutions of a large family of partial differential equations by iteratively minimizing some energy functional plus the Wasserstein distance to the previous iterates. We refer to Section 4.6 for more details and references about these schemes. Following the method introduced in [27], one can approximate these iterations using entropic regularization. Our dual approach is crucial to be able to tackle non-separable energies, such as for instance the total variation of images.

Another illustration of the usefulness of our dual approach is the application to image segmentation developed in [28]. Note that this application requires to compute the gradient of the dual of the smoothed Wasserstein distance with respect to two histograms. This formula is provided in Appendix A.

1.3. Contributions. Our main contribution is to combine the strengths of the dual formulation of [13] with the smoothing strategy laid out by [16] to obtain a *smooth* optimization problem whose objectives and derivatives can be computed in *closed form* in §2. We show that this approach can be readily used to compute Wasserstein barycenters in §3, and explain why using regularized Wasserstein distances might be beneficial to recover smooth solutions. We proceed with more general energies that involve not only Wasserstein distances, but also more generally spatial regularization of barycenters and gradient flows, in §4.

The source code to reproduce the numerical illustration of this article can be found online[1].

1.4. Notations. When used on matrices, functions such as log or exp are always applied element-wise. For two matrices (or vectors) A, B of the same size, $A \circ B$ (resp. A/B) stands for the element-wise product (resp. division) of A by B. If u is a vector, $\mathrm{diag}(u)$ is the diagonal matrix with diagonal u. $\mathbb{1}_n \in \mathbb{R}^n$ is the (column) vector of ones.

[1] 2015-SIIMS-wasserstein-dual

2. Legendre Transforms of the Smoothed Wasserstein Distance.

We introduce in this section the entropic regularization of the Wasserstein distance, study its Legendre transform and show that it admits a simple closed form.

2.1. Optimal Transport with Entropic Smoothing.

We consider two discrete probability distributions on the same space, represented through their histograms $p, q \in \Sigma_n$ of n values. We also introduce a symmetric cost matrix $M = (M_{ij})_{i,j=1,\ldots,n} \in \mathbb{R}_+^{n\times n}$. Each element M_{ij} accounts for the (ground) cost of moving mass from bin i to bin j. In many applications of optimal transport, the cost matrix M is defined through n points $(x_i)_i$ taken in a metric space $(\mathcal{X}, D)$ such that $M_{ij} = D(x_i, x_j)^\rho, \rho \geqslant 1$. Note however that we make no assumption on M in this paper other than the fact that it is symmetric and non-negative.

Given p, q, the set of couplings $U(p,q)$ and the discrete entropy of any coupling in that set are defined as,

$$U(p,q) \overset{\text{def}}{=} \left\{ X \in \mathbb{R}_+^{n\times n} \; ; \; X\mathbb{1}_n = p, X^T\mathbb{1}_n = q \right\}, \; E(X) \quad \overset{\text{def}}{=} -\sum_{ij} h(X_{ij}), \tag{2.1}$$

where $\forall x > 0, h(x) \overset{\text{def}}{=} x \log x, h(0) = 0$. We follow [15]'s approach and introduce a entropy-regularized optimal transport problem:

$$W_\gamma(p,q) \overset{\text{def}}{=} \min_{X \in U(p,q)} \langle M, \, X \rangle - \gamma E(X), \tag{2.2}$$

where $\gamma \geqslant 0$. For $\gamma = 0$, one recovers the usual optimal transport problem, which is a linear program. W_0 is known as the Wasserstein distance (or Earth Mover's Distance, EMD) between p and q. For $\gamma > 0$, Problem (2.2) is strongly convex and admits a unique optimal coupling $X_\gamma^\star$. [15] called the resulting cost $\langle M, \, X_\gamma^\star \rangle$ the *Sinkhorn divergence* between p and q.

While $X_\gamma^\star$ is not necessarily unique for $\gamma = 0$, we show in the following proposition that in the small γ limit, the regularization captures the maximally entropic coupling.

Proposition 2.1. *One has $W_\gamma \to W_0$ as $\gamma \to 0$, and denoting $X_\gamma^\star$ the unique solution of* (2.2)*, one has*

$$X_\gamma^\star \longrightarrow X_0^\star = \underset{X \in U(p,q)}{\operatorname{argmax}} \; \left\{ E(X) \; ; \; \langle M, \, X \rangle = W_0(p,q) \right\}. \tag{2.3}$$

Proof. We consider a sequence $(\gamma_\ell)_\ell$ such that $\gamma_\ell \to 0$ and $\gamma_\ell > 0$. We denote $X_\ell = X_{\gamma_\ell}^\star$. Since $U(p,q)$ is bounded, we can extract a sequence (that we do not relabel for sake of simplicity) such that $X_\ell \to X^\star$. Since $U(p,q)$ is closed, $X^\star \in U(p,q)$. We consider any X such that $\langle M, \, X \rangle = W_0(p,q)$. By optimality of X and X_ℓ for their respective optimization problems (for $\gamma = 0$ and $\gamma = \gamma_\ell$), one has

$$0 \leqslant \langle M, \, X_\ell \rangle - \langle M, \, X \rangle \leqslant \gamma_\ell (E(X_\ell) - E(X)). \tag{2.4}$$

Since E is continuous, taking the limit $\ell \to +\infty$ in this expression shows that $\langle M, \, X^\star \rangle = \langle M, \, X \rangle$ so that $X^\star$ is a feasible point of (2.3). Furthermore, dividing by γ_ℓ in (2.4) and taking the limit shows that $E(X) \leqslant E(X^\star)$, which shows that $X^\star$ is a solution of the maximization (2.3). Since the solution $X_0^\star$ to this program is unique by strict convexity of $-E$, one has $X^\star = X_0^\star$, and the whole sequence is converging. □

[16] provided a dual expression for W_γ. The proof of that result follows from an application of Fenchel-Rockafellar duality to the primal problem (2.2). The indicator function of a closed convex set $\mathcal{C}$ is $\iota_{\mathcal{C}}(x) = 0$ for $x \in \mathcal{C}$ and $\iota_{\mathcal{C}}(x) = +\infty$ otherwise.

Proposition 2.2. *One has*

$$W_\gamma(p, q) = \max_{u,v \in \mathbb{R}^n} \langle u,\, p\rangle + \langle v,\, q\rangle - B(u, v), \tag{2.5}$$

$$B(u, v) \overset{\text{def.}}{=} \begin{cases} \gamma \sum_{i,j} \exp(\frac{1}{\gamma}(u_i + v_j - M_{ij}) - 1), & \text{if } \gamma > 0; \\ \iota_{\mathcal{C}_M}(u, v), \ \text{if } \gamma = 0, \quad \text{where} \quad \mathcal{C}_M \overset{\text{def.}}{=} \{(u, v)\ ;\ u_i + v_j \leqslant M_{ij}\}\,. \end{cases}$$

When $\gamma > 0$, this regularization results in a smoothed approximation of the Wasserstein distance with respect to either of its arguments, as shown below. To simplify notations, let us introduce the notation $H_q(p)$, the Wasserstein distance of any point p to a fixed histogram $q \in \Sigma_n$,

$$\forall\, p \in \Sigma_n, \quad H_q(p) \overset{\text{def}}{=} W_\gamma(p, q).$$

Note that H_q is a convex function for all $\gamma \geqslant 0$. When $\gamma > 0$, H_q has the following properties, which follow from the direct differentiation of expression (2.5):

Proposition 2.3. *For $\gamma > 0$ and $(p, q) \in \Sigma_n \times \Sigma_n$ with $p > 0, q > 0$, H_q is C^1 at p and $\nabla H_q(p) = u^\star$ where $u^\star$ is the unique solution of* (2.5) *satisfying $\langle u^\star,\, \mathbb{1}_n\rangle = 0$.*

Computing both H_q and its gradient requires thus the resolution of the optimization problem in Eq. (2.5), which can be solved with a Sinkhorn fixed-point iteration [15] as remarked by [16, §5]. This computation can be avoided when studying the Fenchel-Legendre conjugate of H_q, as shown below.

2.2. Legendre Transform with Respect to One Histogram.

The goal of this section is to show that the Fenchel-Legendre transform of H_q,

$$\forall\, g \in \mathbb{R}^n, \quad H_q^*(g) = \max_{p \in \Sigma_n} \langle g,\, p\rangle - H_q(p),$$

has a closed form. This result was already known when $\gamma = 0$, that is for the original Wasserstein distance. [13, Prop. 4.1] showed indeed that computing H_q^* only requires a sequence of nearest-neighbor assignments. We show that for $\gamma > 0$, these nearest-neighbor assignments are replaced by soft assignments.

Compared to the primal smoothed Wasserstein distance H_q, the computation of both H_q^* and its derivatives can be carried out without having to solve a matrix-scaling problem. These properties are at the core of the computational framework we develop in this paper.

Theorem 2.4 (Legendre Transform of H_q). *For $\gamma > 0$, the Fenchel-Legendre dual function H_q^* is C^∞. Its gradient function $\nabla H_q^*(\cdot)$ is $1/\gamma$ Lipschitz. Its value, gradient and Hessian at $g \in \mathbb{R}^n$ are, writing $\alpha = e^{g/\gamma}$ and $K = e^{-M/\gamma}$,*

$$\begin{aligned} H_q^*(g) &= \gamma\left(E(q) + \langle q,\, \log K\alpha\rangle\right), \ \nabla H_q^*(g) = \alpha \circ \left(K \frac{q}{K\alpha}\right) \in \Sigma_n, \\ \nabla^2 H_q^*(g) &= \frac{1}{\gamma} \operatorname{diag}\left(\alpha \circ K \frac{q}{K\alpha}\right) - \frac{1}{\gamma} \operatorname{diag}(\alpha) K \operatorname{diag}\left(\frac{q}{(K\alpha)^2}\right) K \operatorname{diag}(\alpha). \end{aligned} \tag{2.6}$$

Proof. Writing $H_{q,M}(p)$ in place of $H_q(p)$ to make explicit the dependency on M, one has

$$\begin{aligned} H^*_{q,M}(g) &= \max_{p\in\Sigma_n} \langle g,\, p\rangle - \max_{u,v} \langle u,\, p\rangle + \langle v,\, q\rangle - B(u,v) \\ &= \max_{p\in\Sigma_n} \; - \max_{u',v} \langle u',\, p\rangle + \langle v,\, q\rangle - B(u'+g,v) \\ &= \max_{p\in\Sigma_n} \; - H_{q,M-g\mathbb{1}^T}(p) = - \min_{p\in\Sigma_n} \min_{X\in U(p,q)} \langle M - g\mathbb{1}^T,\, X\rangle - \gamma E(X). \end{aligned}$$

This leads to an optimal transport problem which is only constrained by *one* marginal,

$$H^*_{q,M}(g) = - \min_{X, X^T\mathbb{1}=q, X\geqslant 0} \langle M - g\mathbb{1}^T,\, X\rangle - \gamma E(X) \tag{2.7}$$

which can be explicitly solved by writing first order conditions for (2.7) to obtain that, at the optimum, we necessarily have $\log(X^\star_{ij}) = \frac{1}{\gamma}(g_i - M_{ij} - \rho_j) - 1$ for some vector of values $\rho \in \mathbb{R}^n$. Therefore $X^\star$ has the form $X^\star = \operatorname{diag}(\alpha) K \operatorname{diag}(e^{\rho/\gamma - 1})$, using the notation $\alpha = e^{g/\gamma}$. Because of the marginal constraint that $X^{\star T}\mathbb{1} = q$, the rightmost diagonal matrix must necessarily be equal to $\operatorname{diag}(q/K\alpha)$, and thus $X^\star = \operatorname{diag}(\alpha) K \operatorname{diag}(q/K\alpha)$. Therefore, the Legendre transform $H^*_{q,M}$ has a closed form,

$$H^*_{q,M}(g) = -\langle M - g\mathbb{1}^T,\, X^\star\rangle + \gamma E(X^\star) \tag{2.8}$$

which can be simplified to

$$H^*_{q,M}(g) = -\gamma \mathbb{1}_d^T \left((K\alpha) \circ h(q/K\alpha)\right).$$

by using the fact that $X^\star = \operatorname{diag}(\alpha) K \operatorname{diag}(q/K\alpha)$. This equation can be simplified further to obtain the expression provided in Eq. (2.6). Using Eq. (2.8), we have that

$$\nabla H^*_{q,M}(g) = X^\star \mathbb{1} = \alpha \circ \left(K \frac{q}{K\alpha}\right).$$

Computations for the Hessian follow directly, and result in the expression given Eq. (2.6). Since the Hessian can be written as the difference of two positive definite matrices, one diagonal and the other equal to the product of a matrix times its transpose, the trace of $\nabla^2 H^*_q(g)$ is upper bounded by the trace of the first term, which is equal to $\frac{1}{\gamma}$ (recall that $\nabla H^*_{q,M}(g)$ is in the simplex), which proves the $\frac{1}{\gamma}$-Lipschitz continuity of the gradient of H^*_q. □

In some settings, such as the Wasserstein propagation framework of [32], the aim is to minimize Wasserstein distances with respect to two variable arguments. We provide the formulation for the corresponding Legendre transform in Theorem A.1 in the Appendix.

2.3. Un-regularized Case.

The result of Theorem 2.4 is derived in the unregularized case (*i.e.* $\gamma = 0$) in [13]. For the sake of comparison, let us now recall this result using our notations. Given a cost matrix $M \in \mathbb{R}^{n\times n}$ and a vector $g \in \mathbb{R}^n$, we introduce for $i \leqslant n$ the set $N_{M,g}(i) = \operatorname{argmin}_k M_{ik} - g_i$. In other words, $N_{M,g}(i)$ is the set of nearest-neighbors of i with respect to the vector of distances M_{ik} offset by $-g_i$.

A map $\sigma_{M,g} : \{1,\ldots,n\} \to \Sigma_n$ is called a nearest-neighbor map if the vector $\sigma_{M,g}(i)$ only has non-zero values on indices in $N_{M,g}(i)$, namely

$$[\sigma_{M,g}(i)]_j \neq 0 \iff j \in N_{M,g}(i).$$

If $N_{M,g}(i)$ is a singleton $\{j\}$ (the minimization $\min_k M_{ik} - g_i$ admits only one optimal solution) then $\sigma_{M,g}(i)$ is necessarily equal to a Dirac histogram δ_j (we call a Dirac histogram a histogram with mass 1 on only one coordinate, of index j in this case). When $N_{M,g}(i)$ has more than one element, ties have to be taken care off, and this can be carried out arbitrarily, for instance by dividing the mass equally among those nearest neighbors, or by only choosing arbitrarily one of them. We can now recall the result of [13]:

Proposition 2.5 (Carlier et al. 2014, Prop. 4.1). *For $\gamma = 0$ and a nearest-neighbor map $\sigma_{M,g}$, the Fenchel-Legendre dual function H_q^* admits the following vector in its sub-differential $\partial H_q^*(g)$ at $g \in \mathbb{R}^n$,*

$$S_q(g) \overset{\text{def}}{=} \sum_{i \leqslant d} q_i \sigma_{M,g}(i) \in \partial H_q^*(g).$$

Note that $S_q(g)$ is in Σ_n. The value of $H_q^*(g)$ is $\langle S_q(g), g\rangle$.

3. Smooth Dual Algorithms For the Wasserstein Barycenter Problem.

In this section, we use the properties of the Legendre transform of the Wasserstein distance as detailed in Section §2 to solve the Wasserstein Barycenter Problem.

3.1. Smooth Dual Formulation of the WBP.

Following the introduction of the Wasserstein Barycenter Problem (WBP) by [2], [16] introduced the smoothed WBP with γ-entropic regularization (γ-sWBP) as

$$\min_{p \in \Sigma_n} \sum_{k=1}^{N} \lambda_k H_{q_k}(p) \quad . \tag{3.1}$$

where $(q_1, \dots, q_N)$ is a family of histograms in Σ_n. When $\gamma = 0$, the γ-sWBP is exactly the WBP. In that case, problem (3.1) is in fact a linear program, as discussed later in §3.4. When $\gamma > 0$ the γ-sWBP is a *strictly* convex optimization problem that admits a unique solution, which can be solved with a simple gradient descent as advocated by [16]. They show that the N gradients $[\nabla H_{q_k}(p)]_{k \leqslant N}$ can be computed at each iteration by solving N Sinkhorn matrix-scaling problems. Because these gradients are themselves the result of a numerical optimization procedure, the problem of choosing an adequate threshold to obtain sufficiently precise gradients arises as a key parameter in that approach. We take here a different route to solve the γ-sWBP, which can be either interpreted as a smooth alternative to the dual WBP studied by [13], or the dual counterpart to the smoothed WBP of [16].

Theorem 3.1. *The barycenter $p^\star$ solving (3.1) satisfies*

$$\forall\, k = 1, \dots, N, \quad p^\star = \nabla H_{q_k}^*(g_k^\star) \tag{3.2}$$

where $(g_k^\star)_k$ are any solution of the smoothed dual WBP:

$$\min_{g_1, \dots, g_N \in \mathbb{R}^n} \sum_k \lambda_k H_{q_k}^*(g_k) \quad s.t. \quad \sum_k \lambda_k g_k = 0. \tag{3.3}$$

Proof. We re-write the barycenter problem

$$\min_{p_1, \dots, p_N} \sum_k \lambda_k H_{q_k}(p_k) \quad \text{s.t.} \quad p_1 = \dots = p_N$$

whose Fenchel-Rockafelar dual reads

$$\min_{\tilde g_1,\dots,\tilde g_N} \sum_k \lambda_k H^*_{q_k}(\tilde g_k/\lambda_k) \quad \text{s.t.} \quad \sum_k \tilde g_k = 0.$$

Since the primal problem is strictly convex, the primal-dual relationships show that the unique solution $p^\star$ of the primal can be obtained from any solution $(g_k^\star)_k$ via the relation $p_k^\star = \nabla H^*_{q_k^\star}(\tilde g_k^\star/\lambda_k)$. One obtains the desired formulation using the change of variable $g_k = \tilde g_k/\lambda_k$. □

Theorem 3.1 provides a simple approach to solve the γ-sWBP: Rather than minimizing directly the sum of regularized Wasserstein distances in Eq. (3.1), this formulation only involves minimizing a strictly convex function with closed form objectives and gradients.

Parallel Implementation. The objectives, gradients and Hessians of the Fenchel-Legendre dual H_q^* can be computed using either matrix-vector products or element-wise operations. Given N histograms $(q_k)_k$, N dual variables $(g_k)_k$ and N arbitrary vectors $(x_k)_k$, the computation of N objective values $(H^*_{q_k}(g_k))_k$ and N gradients $(\nabla H^*_{q_k}(g_k))_k$ can all be vectorized. Assuming that all column vectors g_k, q_k and x_k are gathered in $n \times N$ matrices G, Q and X respectively, we define first the following $n \times N$ auxiliary matrices:

$$A \overset{\text{def}}{=} e^{G/\gamma}, \quad B \overset{\text{def}}{=} KA, \quad C \overset{\text{def}}{=} \frac{Q}{B}, \quad \Delta \overset{\text{def}}{=} A \circ (KC),$$

to form the vector of objectives

$$H^* \overset{\text{def}}{=} [H^*_{q_1}(g_1), \dots, H^*_{q_N}(g_N)] = -\gamma \mathbb{1}_n^T \left(Q \circ \log(C)\right),$$

and the matrix of gradients

$$\nabla H^* \overset{\text{def}}{=} [\nabla H^*_{q_1}(g_1), \dots, \nabla H^*_{q_N}(g_N)] = \Delta. \tag{3.4}$$

3.2. Algorithm. The γ-sWBP in Eq. (3.3) has a smooth objective with respect to each of its variables g_k, a simple linear equality constraint, and both gradients and Hessians that can computed in closed form. We can thus compute a minimizer for that problem using a naive gradient descent outlined in Algorithm 1. To obtain a faster convergence, it is also possible to use accelerated gradient descent, quasi-Newton or truncated Newton methods [11, §10]. In the latter case, the resulting KKT linear system is sparse, and solving it with preconjugate gradient techniques can be efficiently carried out. We omit these details and only report results using off-the-shelf L-BFGS. From the dual iterates g_k stored in a $n \times N$ matrix G, one recovers primal iterates using the formula (3.2), namely $p_k = e^{g_k/\gamma} \circ K\frac{q_k}{Ke^{g_k/\gamma}}$. At each intermediary iteration one can thus form a solution to the smoothed Wasserstein barycenter problem by averaging these primal solutions,$\tilde p = \Delta \mathbb{1}_N/N$. Upon convergence, these p_k are all equal to the unique solution $p^\star$. The average at each iteration $\tilde p$ converges towards that unique solution, and we use the sum of all line wise standard deviations of Δ: $\mathbb{1}_d^T\sqrt{(\tilde\Delta \circ \tilde\Delta)\mathbb{1}_N/N}$ where $\tilde\Delta = \Delta(I_N - \frac{1}{N}\mathbb{1}_N\mathbb{1}_N^T)$ to monitor that convergence in our algorithms.

3.3. Initialization Heuristic. Definition 3.2 provides an initialization heuristic to initialize both the primal and dual smoothed WBP, motivated by the fact that they

Algorithm 1 Smoothed Wasserstein Barycenter, Generic Algorithm

```
Input: $Q = [q_1, \cdots, q_N] \in (\Sigma_n)^N$, metric $M \in \mathbb{R}_+^{n\times n}$, barycenter weights $\lambda \in \Sigma_N$, $\gamma > 0$, tolerance $\varepsilon > 0$.
initialize $G \in \mathbb{R}^{n\times N}$ and form the $n \times n$ matrix $K = e^{-M/\gamma}$.
repeat
    From gradient matrix $\Delta$ (see Eq. 3.4) produce update matrix $\hat{\Delta}$ using either $\Delta$ directly or other methods such as L-BFGS.
    $G = G - \tau\hat{\Delta}$, update with fixed step length $\tau$ or approximate line search to set $\tau$.
    $G = G - \frac{1}{\|\lambda\|_2^2}(G\lambda)\lambda^T$     (projection such that $G\lambda = 0$)
until $\mathbb{1}_d^T\sqrt{(\tilde{\Delta}\circ\tilde{\Delta})\mathbb{1}_N/N} < \varepsilon$, where $\tilde{\Delta} = \Delta(I_N - \frac{1}{N}\mathbb{1}_N\mathbb{1}_N^T)$
output barycenter $p = \Delta\mathbb{1}_N/N$.
```

provide directly the optimal primal/dual solutions when the histograms are Dirac histograms as proved in Proposition 3.3.

Definition 3.2 (Primal and Dual WBP Initialization). *Let $(q_1, \cdots, q_N)$ be N target histograms in the simplex Σ_n and λ a vector of weights in Σ_N. Let $\bar{q} = \sum_k \lambda_k q_k \in \Sigma_n$. Define κ_γ as*

$$\kappa_\gamma = \begin{cases} e^{-M\bar{q}/\gamma}/(\mathbb{1}_n^T e^{-M\bar{q}/\gamma}) & \text{if } \gamma > 0, \\ \delta_j, \ \text{where } j \in \operatorname{argmin}_\ell [M\bar{q}]_\ell, & \text{if } \gamma = 0. \end{cases}$$

For $\gamma \geqslant 0$, the γ-smoothed primal and dual WBP can be initialized respectively with the following primal and N dual feasible solutions:

$$p^{(0)} \overset{\text{def}}{=} \kappa_\gamma, \tag{3.5}$$

and for $1 \leqslant k \leqslant N$,

$$g_k^{(0)} \overset{\text{def}}{=} M(q_k - \bar{q}). \tag{3.6}$$

The primal initialization described above differs when $\gamma > 0$ or $\gamma = 0$: For $\gamma > 0$, κ_γ is the normalized, weighted geometric average of the columns of the kernel $K = e^{-M/\gamma}$; when $\gamma = 0$, κ_γ is a vector of zero values except for a value of 1 on the index corresponding to the (or any, if many) smallest entry of $M\bar{q}$. On the other hand, the dual initialization is the same for both smoothed and non-smoothed Wasserstein barycenter problems.

The initializations proposed in Definition 3.2 solve the WBP in the case that all histograms are Dirac histograms, as proved in Proposition 3.3. For more general problems, we have observed that this initialization is particularly useful when solving the WBP with the dual formulation, but not so much with the primal one. In many experimental problems we have considered, the dual initialization seems to capture important features of the optimal solution. The primal solution that results from this dual initialization, that obtained by averaging the gradients $\nabla H_{q_k}^*(g_k^\star)$ as suggested by the primal/dual relation of Equation (3.1), can serve as a rough approximation of the barycenter. We provide its explicit expression $p_{\text{dual}}^{(0)}$ below. Note that $p_{\text{dual}}^{(0)}$ differs from the initialization $p^{(0)}$ suggested in Equation (3.5).

$$p_{\text{dual}}^{(0)} = \frac{1}{n}\left(e^{M(Q-\frac{1}{n}Q\mathbb{1}_n\mathbb{1}_N^T)/\gamma} \circ \left(K\frac{Q}{Ke^{M(Q-\frac{1}{n}Q\mathbb{1}_n\mathbb{1}_N^T)}}\right)\right)\mathbb{1}_n.$$

Proposition 3.3. *Let λ be a vector of weights in Σ_N, and $(q_1, \cdots, q_N)$ be N Dirac histograms, namely histograms that are zero everywhere but for one coordinate equal to 1. For $\gamma \geqslant 0$, the γ-sWBP primal and dual problems are solved exactly using the initialization of Definition* (3.2).

Proof. To simplify notations, we write $p = p^{(0)}$ and $g_k = g_k^{(0)}$ as defined in Definition 3.2 above. First, one can easily check that both initialization satisfy the necessary constraints, *i.e.* $p \in \Sigma_n$ and $\sum_k \lambda_k g_k = 0$.

When $\gamma = 0$, since all q_k are Dirac histograms, the Wasserstein distance of any point x in the simplex to any q_k is equal to $x^T M q_k$. Therefore, the Wasserstein barycenter objective evaluated at x is equal to $x^T M\bar{q}$. This can be trivially minimized by selecting any histogram giving a mass of 1 to the index corresponding to any smallest entry in the vector $M\bar{q}$, which is the definition of p. A similar computation for the dual problem results in the dual optimal outlined above.

When $\gamma > 0$, we need to prove that each gradient of $H_{q_k}^*$ computed at g_k is equal to p for all $1 \leqslant k \leqslant N$. Writing $\alpha_k = e^{g_k/\gamma}$, we recover that

$$\alpha_k = \frac{\kappa_\gamma}{\xi_k},$$

where $\xi_k \overset{\text{def}}{=} e^{-Mq_k/\gamma}$. Since q_k is a Dirac histogram, all of its coordinates are equal to 0, but for one coordinate whose value is 1. Let j be the index of that coordinate. Therefore, $\xi_k \overset{\text{def}}{=} e^{-Mq_k/\gamma} = K_j$, where K_j is the j^{th} column of the matrix $K = e^{-M/\gamma}$. Therefore,

$$\alpha_k = \frac{\kappa_\gamma}{K_j}.$$

Let us now compute the gradient ∇_k of $H_{q_k}^*$ at g_k by following Eq. (2.6):

$$\nabla_k = \alpha_k \circ \left(K \frac{q_k}{K\alpha_k} \right).$$

Because of the symmetry of K, we have that the j^{th} element of the vector $K\alpha_k$ is equal to:

$$(K\alpha_k)_j = K_j^T \alpha_k = \mathbb{1}_n^T (K_j \circ \alpha_k) = \mathbb{1}_n^T \left(K_j \circ \left(\frac{\kappa_\gamma}{K_j} \right) \right) = 1.$$

Since only the j^{th} element of q_k is non-zero by definition, $q_k/(K\alpha_k) = q_k$. Because q_k is everywhere zero except for its j^{th} coordinate, $K(q_k/K\alpha_k)$ is thus equal to the j^{th} column of K, namely

$$K \frac{q_k}{K\alpha_k} = K_j.$$

Finally, we obtain that the gradient of $H_{q_k}^*$ at g_k is equal to

$$\nabla_k = \alpha_k \circ \left(K \frac{q_k}{K\alpha_k} \right) = \frac{\kappa_\gamma}{K_j} \circ K_j = \kappa_\gamma = p^{(0)},$$

which holds for all indices $1 \leqslant k \leqslant N$. □

3.4. Smoothing and Stabilization of the WBP. We make the claim in this section that smoothing the WBP is not only beneficial computationally, but may also yield more stable computations. Of central importance in this discussion is the fact that the WBP can be cast as a LP of $Nn^2 + n$ variables and $2Nn$ constraints, and thus solved *exactly* for small n and N:

$$\begin{aligned} \min_{X_1,\cdots,X_N,p} & \sum_{k=1}^{N} \lambda_k \langle X_k,\, M\rangle \\ \text{s.t. } & X_k \in \mathbb{R}_+^{n\times n}, \forall k \leqslant N; p \in \Sigma_n, \\ & X_k^T \mathbb{1}_n = q_k, \forall k \leqslant N, \\ & X_1\mathbb{1}_n = \cdots = X_N \mathbb{1}_n = p. \end{aligned} \tag{3.7}$$

Given couplings $X_1^\star, \cdots, X_N^\star$ which are optimal solutions to Eq.(3.7), the solution to the WBP is equal to the marginal common to all those couplings: $p^\star = X_k^\star \mathbb{1}_n$ for any $k \leqslant N$. For small N and n, this problem is tractable, but it can be surprisingly ill-posed as we see next.

Indeed, it is also known that the 2-Wasserstein mean of two univariate (continuous) Gaussian densities of mean and standard deviation (μ_1, σ_1) and (μ_2, σ_2) respectively is a Gaussian of mean $(\mu_1 + \mu_2)/2$ and standard deviation $(\sigma_1 + \sigma_2)/2$ [2, §6.3]. This fact is illustrated in the top-left plot of Figure 3.1 where we display the average Wasserstein average $\mathcal{N}(0, 5/8)$ of the two densities $\mathcal{N}(2, 1)$ and $\mathcal{N}(-2, 1/4)$. That plot is obtained by using smoothed spline interpolations of a uniformly spaced grid of 100 values, as can be better observed in the top-right (stair) plot, where the discrete evaluations of these densities are respectively denoted p_W, q_1 and q_2.

Naturally, one would expect the barycenter of q_1 and q_2 to be close, in some sense, to the discretized histogram p_W of their true barycenter. Histogram $p^\star$, displayed in the bottom-left plot, is the exact optimal solution of Eq. (3.7), computed with the simplex method. That WBP reduces to a linear program of 2×100^2 variables and 300 constraints. We observe that $W_2^2(p^\star, q_1) + W_2^2(p^\star, q_2) = 0.5833950$ whereas $W_2^2(p_W, q_1) + W_2^2(p_W, q_2) = 0.5834070$. The solution obtained with the simplex has, indeed, a smaller objective than the discretized version of the true barycenter.

The bottom-right plot displays the solution of the *smoothed* Wasserstein barycenter problem (with smoothing parameter $\gamma = \frac{1}{100}$ and a ground cost M that has been re-scaled to have a median value of 1). The objective value for that smoothed approximation is 0.5834597.

This numerical experiment does not contradict the fact that the discretized barycenter $p^\star$ converges to the continuous barycenter as the grid size tends to zero, as shown in [13]. This observation illustrates however that, because it is defined as the argmin of a linear program, the true Wasserstein barycenter may be extremely unstable, even for such a simple problem and for large n as illustrated in Figure 3.2. Regularizing the Wasserstein distances has thus the added benefit of smoothing the resulting solution of the Wasserstein barycenter problem, and that of mitigating low sample size effects.

3.5. Performance on the Wasserstein Barycenter Problem. We compare in this section the behavior of the smooth dual approach presented in this paper with that of (i) the smooth primal approach of [16], (ii) the dual approach of [13], and (iii) the Bregman iterative projections approach of [7]. We compare these methods on the simple task of computing the Wasserstein barycenter of 12 histograms laid out on the 100×100 grid, as previously introduced in [7, §3.2]. We outline briefly all four methods below, and follow by presenting numerical results.

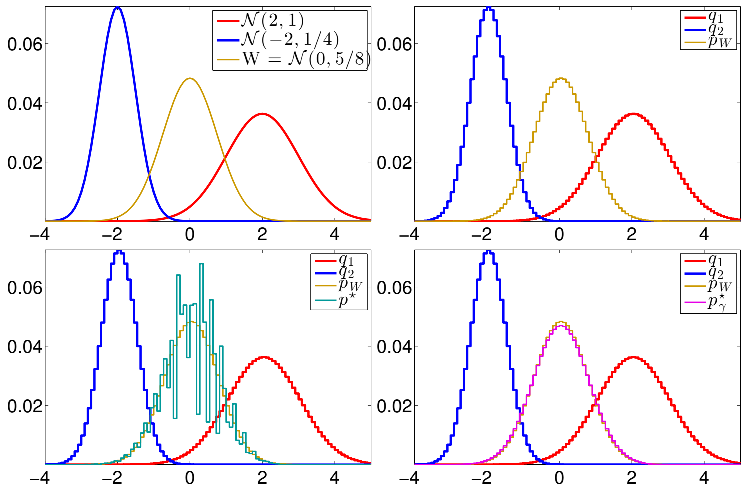


Fig. 3.1: (top-left) two Gaussian densities and their barycenter (top right) same densities, discretized (bottom left) discretization of the true barycenter *vs.* the optimum of Equation 3.7 (bottom right) barycenter computed with our smoothing approach.

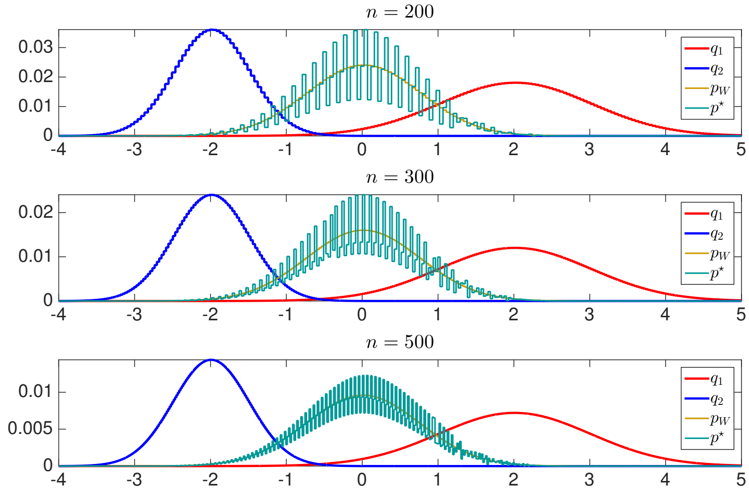


Fig. 3.2: Plots of the exact barycenters for varying grid size n.

Smooth primal first-order descent. [16, §5] proposed to minimize directly Eq. (3.1) with a regularizer $\gamma > 0$. That objective can be evaluated by running N Sinkhorn fixed-point iterations in parallel. That objective is differentiable and its gradient is equal to $\gamma \sum_k \lambda_k \log \alpha_k$, where the α_k are the left scalings obtained with that subroutine. A weakness of that approach is that a tolerance ε for the Sinkhorn fixed-point algorithm must be chosen. Convergence for the Sinkhorn algorithm can be measured with a difference in l_1 norm (or any other norm) between the row and column marginals of $\operatorname{diag}(\alpha_k)e^{-M/\gamma}\operatorname{diag}(\beta_k)$ and the targeted histograms p and q_k. Setting that tolerance ε to a large value ensures a faster convergence of the subroutine, but would result in noisy gradients which could slow the convergence of the algorithm. Because the smoothed dual approach only relies on closed form expressions we dot

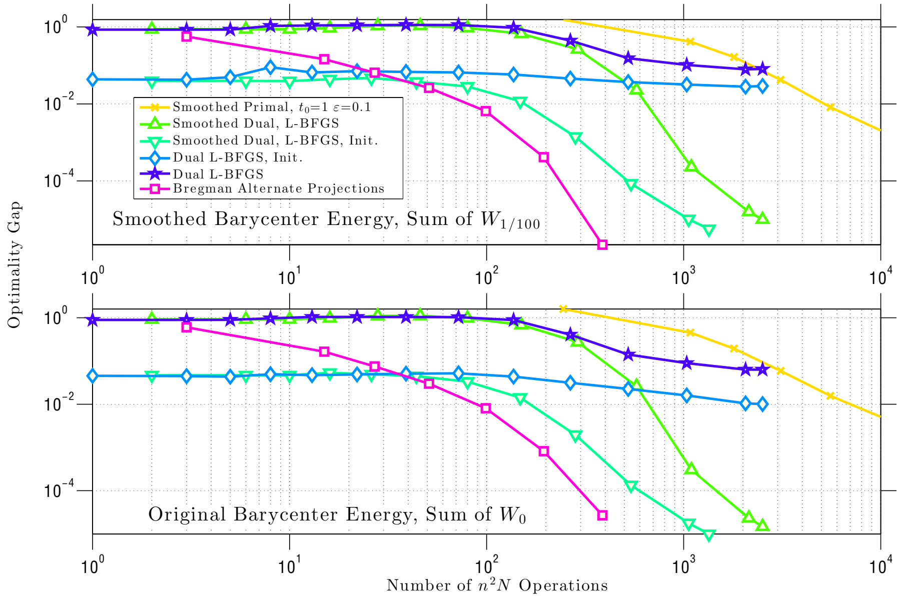


Fig. 3.3: Number of quadratic operations (matrix vector product or min search in a matrix) *vs.* optimization gap to the smallest possible objective found after running 10^5 iterations of all algorithms, *log-log* scale. Because the smooth primal/dual approaches optimize a *different* criterion than the dual approach, we plot *both* objectives. The Smooth dual L-BFGS converges faster in *both* smooth and non-smooth metrics. Note the crucial importance of the initialization proposed in §3.3.

not have to take into account such a trade-off.

Iterative Bregman Projections. [7, Prop. 1] recalls that the computation of the smoothed Wasserstein distance between p, q using the Sinkhorn algorithm can be interpreted as an iterative alternated projection of the $n \times n$ kernel matrix $e^{-M/\gamma}$ onto two affine sets, $\{X : X\mathbb{1}_n = p\}$ and $\{X : X^T\mathbb{1}_n = q\}$. That projection is understood to be in the Kullack-Leibler divergence sense. More interestingly, the authors also show that the smoothed WBP itself can also be tackled using an iterative alternated projection, cast this time in a space of dimension $n \times n \times N$. Very much like the original Sinkhorn algorithm, these projections can be computed for a cheap price, by only tracking variables of size $n \times N$. This approach yields an extremely simple, parameter-free, generalization of the Sinkhorn algorithm which can be applied to the WBP.

Smooth dual L-BFGS. The dual formulation with variables $(g_1, \cdots, g_N) \in (\mathbb{R}^n)^N$ of Eq. (3.3) can be solved using a constrained L-BFGS solver At each iteration of that minimization, we can recover a feasible solution p to the primal problem of Eq. (3.1) via the primal-dual relation $p = \frac{1}{N}\sum_k \nabla H^*_{q_k}(\tilde{g}_k)$.

Dual ($\gamma = 0$) *with L-BFGS.* This approach amounts to solving directly the (non-differentiable) dual problem described in Eq. (3.3) with no regularization, namely $\gamma = 0$. Subgradients for the Fenchel-Legendre transforms $H^*_{q_k}$ can be obtained in closed form through Proposition 2.5. As with the smoothed-dual formulation, we can also obtain a feasible primal solution by averaging subgradients. We follow [13]'s recommendation to use L-BFGS. The non-smoothness of that energy is challenging: We have observed empirically that a naive subgradient method applied to that problem fails to converge in all examples we have considered, whereas the L-BFGS approach converges, albeit without guarantees.

Averaging Truncated Mixtures of Gaussians. We consider the 12 truncated mixtures of Gaussians introduced in [7, §3.2]. To compare computational time, we use Nn^2 elementary operations as the computation unit. These Nn^2 operations correspond to matrix-matrix products in the smoothed Wasserstein case, and Nn computations of nearest neighbor assignments among n possible neighbors. Note that in both cases (Gaussian matrix product and nearest neighbors under the L_2 metric) computations can be accelerated by considering fast Gaussian convolutions and *kd*-trees for fast nearest neighbor search. We do not consider them in this section. We plot the optimality gap w.r.t the optimum of these 4 techniques as a function of the number of computations, by taking as a reference the lowest value attained across all methods. This value is attained, as in [7], by the iterative Bregman projections approach after 771 iterations (not displayed in our graph). We show these gaps for both the smoothed ($\gamma = 1/100$) and non-smoothed objectives ($\gamma = 0$). We observe that the iterative Bregman approach outperforms all other techniques. The smoothed-dual approach follows closely, notably when initialized with the formula provided in Definition 3.2.

4. Regularized Problems.

We show in this section that our dual optimization framework is versatile enough to deal with functionals involving Wasserstein distances that are more general than the initial WBP problem.

4.1. Regularized Wasserstein Barycenters.

In order to enforce additional properties of the barycenters, it is possible to penalize (3.1) with an additional convex regularization, and consider

$$\min_p \sum_{k=1}^N w_k H_{q_k}(p) + J(\mathcal{A}p), \tag{4.1}$$

where J is a convex real-valued function, and $\mathcal{A}$ is a linear operator.

The following proposition shows how to compute such a regularized barycenter through a dual optimization problem.

Proposition 1. *The dual problem to (4.1) reads*

$$\min_{(u_k)_{k=1}^N, v} \sum_{k=1}^N w_i H^*_{q_k}(u_k) + J^*(v) + \iota_H((u_k)_k, v) \tag{4.2}$$

$$\textit{where} \quad H \overset{\text{def.}}{=} \left\{ ((u_k)_{k=1}^N, v) \;;\; \mathcal{A}^* v + \sum_k w_k u_k = 0 \right\},$$

and the primal-dual relationships read

$$\forall\, k = 1, \ldots, N, \quad p = \nabla H^*_{q_k}(u_k). \tag{4.3}$$

Proof. We re-write the initial program (4.1) as

$$\min_\pi F(B\pi) + G(\pi) \tag{4.4}$$

where we denoted, for $\pi = (p, p_1, \ldots, p_N)$,

$$\begin{aligned} B\pi &\overset{\text{def.}}{=} (\mathcal{A}p, p, p_1, \ldots, p_N) \\ F(\beta, q, p_1, \ldots, p_N) &\overset{\text{def.}}{=} J(\beta) + \iota_C(q, p_1, \ldots, p_N), \\ G(p, p_1, \ldots, p_N) &\overset{\text{def.}}{=} \sum_k w_k H_{q_k}(p_k) \end{aligned}$$

for $C \overset{\text{def.}}{=} \{(q, p_1, \ldots, p_N) \; ; \; \forall k, p_k = q\}$. The Fenchel-Rockafelear dual to (4.4) reads

$$\max_{\nu = \{v, u, (u_k)_k\}} \quad -F^*(\nu) - G^*(-B^*\nu)$$

where

$$\begin{aligned} G^*(u, u_1, \ldots, u_N) &= \sum_k w_k H^*_{q_k}(u_k/w_k) + \iota_{\{0\}}(u), \\ B^*(\nu) &= (\mathcal{A}^* v + u, u_1, \ldots, u_N), \\ F^*(\nu) &= J^*(v) + \iota_{C^\perp}(u, u_1, \ldots, u_N), \end{aligned}$$

where $C^\perp = \{(u, u_1, \ldots, u_N) \; ; \; u + \sum_k u_k = 0\}$. One thus obtains the dual

$$\min_{v, b, (u_k)_k} \sum_k w_k H^*_{q_k}(-u_k/w_k) \quad \text{s.t.} \quad \left\{ \begin{array}{l} \mathcal{A}^* v + u = 0, \\ u + \sum_k u_k = 0. \end{array} \right.$$

The primal-dual relationships reads $\pi \in \partial G^*(-B^*\nu)$, and hence (4.3). Changing $-u_k/w_k$ into u_k give the desired formula. □

Relevant examples of penalizations J include:

- In order to enforce some spread of the barycenter, one can use $\mathcal{A} = \text{Id}$ and $J(p) = \frac{\lambda}{2}\|p\|^2$, in which case $J^*(g) = \frac{1}{2\lambda}\|g\|^2$. In contrast to (3.3), the dual problem (4.2) is equivalent to an unconstraint smooth optimization. This problem can be solved using a simple Newton descent.
- One can also enforce that the barycenter entries are smaller than some maximum value ρ by setting $\mathcal{A} = \text{Id}$ and $J = \iota_{\mathcal{C}}$ where $\mathcal{C} = \{p \; ; \; \|p\|_\infty \leqslant \rho\}$. In this case, one has $J^*(g) = \rho\|g\|_1$. The optimization (4.2) is equivalent an unconstrained non-smooth optimization. Since the penalization is an ℓ^1 norm, one solve it using first order proximal methods as detailed in Section 4.2 bellow.
- To force the barycenter to assume some fixed values $p_I^0 \in \mathbb{R}^{|I|}$ on a given set I of indices, one can use $\mathcal{A} = \text{Id}$ and $J = \iota_{\mathcal{C}}$ where $\mathcal{C} = \left\{p \; ; \; p_I = p_I^0\right\}$ where $p_I = (p_i)_{i \in I}$. One then has $J^*(g) = \langle g_I, p_I^0 \rangle + \iota_{\{0\}}(g_{I^c})$.
- To force the barycenter to have some smoothness, one can select $\mathcal{A}$ to be a spacial derivative operator (for instance a gradient approximated on some grid or mesh) and J to be a norm such as an ℓ^2 norm (to ensure uniform smoothness) or an ℓ^1 norm (to ensure piecewise regularity). We explore this idea in Section 4.3.

4.2. Resolution using First Order Proximal Splitting. Assuming without loss of generality that $w_N \neq 0$ (otherwise one simply needs to permute the ordering of the input densities), one can note that it is possible to remove u_N from (4.2) by imposing, for $x = ((u_k)_{k=1}^{N-1}, v)$

$$u_N(x) \overset{\text{def.}}{=} -\frac{\mathcal{A}^* v}{w_N} - \sum_{i=1}^{N-1} \frac{w_k}{w_N} u_k,$$

and then one can consider the following optimization problem without the H constraint

$$\min_x F(x) + G(x) \quad \text{where} \quad \left\{ \begin{array}{l} F(x) \overset{\text{def.}}{=} \sum_{k=1}^{N-1} w_i H^*_{q_k}(u_k) + w_N H^*_{q_N}(u_N(x)) \\ G(x) \overset{\text{def.}}{=} J^*(v). \end{array} \right. \tag{4.5}$$

We assume that one is able to compute the proximal operator of J^*

$$\operatorname{Prox}_{\tau J^*}(v) \overset{\text{def.}}{=} \underset{v'}{\operatorname{argmin}} \frac{1}{2}\|v - v'\|^2 + \tau J^*(v'). \tag{4.6}$$

It is for instance an orthogonal projector on a convex set C when $J^* = \iota_C$ is the indicator of C. One can compute easily this projection for instance when J is the ℓ^2 or the ℓ^1 norm (see Section 4.3). We refer to [5] for more background on proximal operators.

The proximal operator of G is then simply

$$\forall\, x = ((u_k)_{k=1}^{N-1}, v), \quad \operatorname{Prox}_{\tau G}(x) = ((u_k)_{k=1}^{N-1}, \operatorname{Prox}_{\tau J^*}(v)).$$

Note also that the function F is smooth with a Lipschitz gradient, and that

$$\nabla F((u_k)_{k=1}^{N-1}, v) = \big((w_k(\nabla H^*_{q_k}(u_k) - \nabla H^*_{q_N}(u_N)))_{k=1}^{N-1}, -\mathcal{A}\nabla H^*_{q_N}(u_N)\big)$$

The simplest algorithm to solve (4.5) is the Forward-Backward algorithm, whose iteration read

$$x^{(\ell+1)} = \operatorname{Prox}_{\tau J^*}\left(x^{(\ell)} - \tau \nabla F(x^{(\ell)})\right). \tag{4.7}$$

It $\tau < 2/L$ where L is the Lipschitz constant of ∇F, then $x^{(\ell)}$ converge to a solution of (4.5), see [5] and the references therein. In order to accelerate the convergence of the method, one can use accelerated schemes such as FISTA's algorithm [6].

4.3. Total Variation Regularization. A typical example of regularization to enforce some geometrical regularity in the barycenter is the total variation regularization on a grid in $\mathbb{R}^d$ (e.g. $d = 2$ for images). It is obtained by considering

$$\mathcal{A}p \overset{\text{def.}}{=} \nabla p = (\nabla_i p)_i \quad \text{and} \quad J(u) \overset{\text{def.}}{=} \lambda \sum_i \|u_i\|_\beta, \tag{4.8}$$

where $\nabla_i p \in \mathbb{R}^d$ is a finite difference approximation of the gradient at a point indexed by i, and $\lambda \geqslant 0$ is the regularization strength. When using the ℓ^2 norm to measure the gradient amplitude, i.e. $\beta = 2$, one obtains the so-called isotropic total variation, that tends to round corners, and essentially penalizes the length of the level sets of the barycenter, possibly merging clusters together. When using instead the ℓ^1 norm, i.e. $\beta = 1$, one obtains the so-called anisotropic total variation, which penalizes independently horizontal and vertical derivative, thus favoring the emergence of axis-aligned edges, and giving a "crystalline" look to the barycenters. We refer for instance to [14] for a study of the effect of TV regularization on the shapes of levelsets using isotropic and crystalline total variations.

In this case, it is possible to compute in closed form the proximal operator (4.6). Indeed, one has $J^* = \iota_{\|\cdot\|_{\beta^*} \leqslant \lambda}$ where β^* is the conjugate exponent $1/\beta + 1/\beta^* = 1$. One can compute explicitly the proximal operator in the case $\beta \in \{1, 2\}$ since they correspond to orthogonal projectors on ℓ^{β^*} balls

$$\operatorname{Prox}_{\tau J^*}(v)_i = \left\{ \begin{array}{ll} \min(\max(v_i, -\lambda), \lambda) & \text{if} \quad \beta = 1, \\ v_i \frac{\lambda}{\max(\|v_i\|, 1)} & \text{if} \quad \beta = 2. \end{array} \right.$$

4.4. Barycenters of Images. We start by computing barycenters of a small number of 2-D images, that are discretized on an uniform rectangular grid of $n = 256 \times 256$ pixels $(z_i)_{i=1}^N$. We use either the isotropic ($\beta = 2$) or anisotropic ($\beta = 1$) total variation presented above, where ∇ is defined using standard forward finite differences along each axis, and using Neumann boundary conditions. The metric is the usual squared Euclidean metric

$$\forall\, (i,j) \in \{1, \ldots, n\}^2, \quad M_{i,j} = \|z_i - z_j\|^2. \tag{4.9}$$

The Gibbs kernel $K = e^{-M/\gamma}$ is a filtering with a Gaussian kernel, that can be applied efficiently to histograms in nearly linear time, see [31] for more details about convolutional kernels.

Figure 4.1 shows examples of barycenters of $N = 4$ input histograms computed by solving (4.1) using the projected gradient descent method (4.7). The input histograms represent 2-D shapes, and are uniform (constant) distributions inside the support of the shapes. Note that in general the barycenters are not shapes, i.e. they are not uniform distributions, but this method can nevertheless be used to define meaningful averaging of shapes as exposed in [31]. Figure 4.1 compares the effects of $\beta \in \{1, 2\}$, and one can clearly see how the isotropic total variation ($\beta = 2$) rounds the corners of the input densities, while the anisotropic version ($\beta = 1$) favors horizontal/vertical edges.

Figure 4.2 shows the influence of the regularization strength λ to compute the iso-barycenter of $N = 2$ shapes. This highlights the fact that this total variation regularization has the tendency to group together small clusters, which might be beneficial for some applications, as illustrated in Section 4.5 on MEG data denoising.

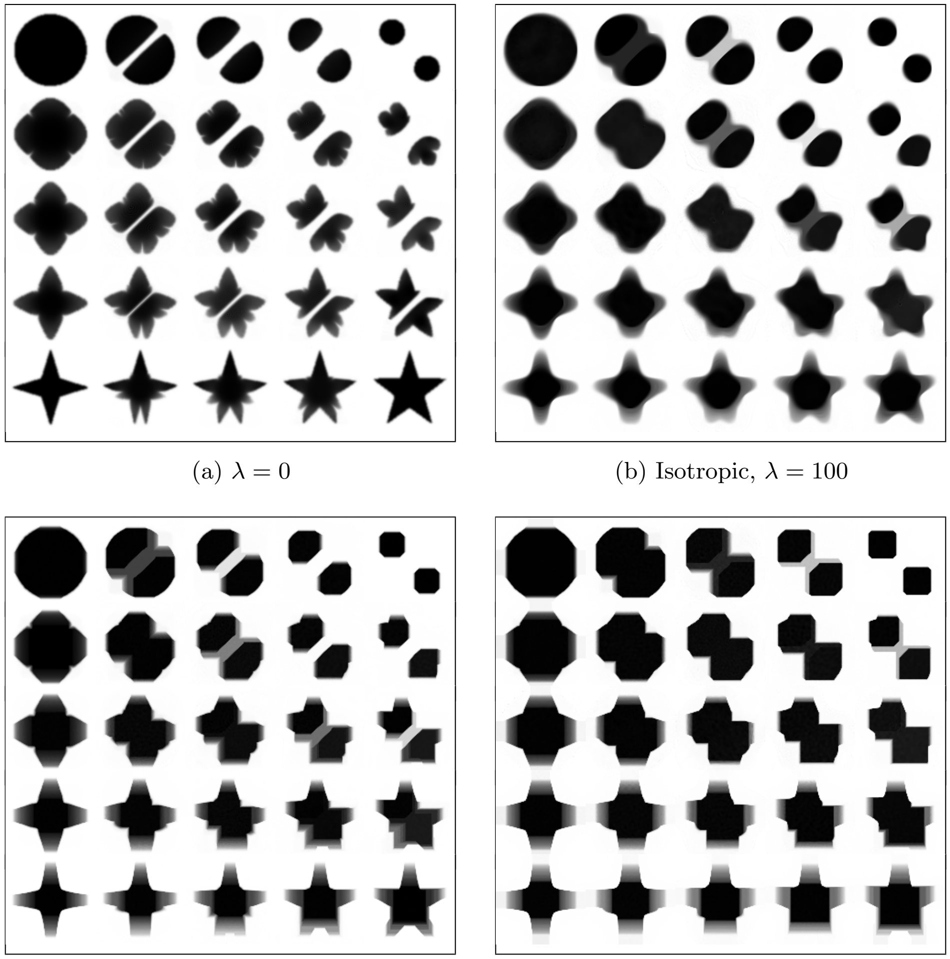
4.5. Barycenters of MEG Data. We applied our method to a magnetoencephalography (MEG) dataset. In this setup, brain activity of a subject is recorded (Elekta Neuromag, 306 sensors of which 204 planar gradiometers and 102 magnetometers, sampling frequency 1000Hz) while the subject reacted to the presentation of a target stimulus by pressing either the left or the right button.

Data is preprocessed applying signal space separation correction, interpolation of noisy sensors, and realignment of data into a subject-specific head position (MaxFilter, Elekta Neuromag). The signal was then filtered (low pass 40HZ), and artifacts such as blinks and heartbeats removed thanks to Signal-Space Projection using the Brainstorm software[2]. The samples we used for our barycenter computations are an average of the norm of the two gradiometers for each channel from stimulation onto 50ms and the classes were left or right button.

This results in two classes of recordings, one for each pressed button. We aim at computing a representative activity map for each class using Wasserstein barycenters. For each class we have $N = 33$ recordings $(q_k)_{k=1}^N$ each having $n = 66$ samples located on the vertices of an hexahedral mesh of a hemisphere (corresponding to a MEG recording helmet). These recorded values are positive by construction, and we rescale them linearly to impose $q_k \in \Sigma_n$. Figure 4.3, top row, shows some samples from this dataset, displayed using interpolated colors as well as iso-level curves. The black dots represent the position $(z_i)_{i=1}^n$ of the electrodes on the half-sphere of the helmet, flattened on a 2-D disk.

We computed TV-regularized barycenters independently for each class by solving (4.1) with the TV regularization using the projected gradient descent method (4.7).

[2] http://neuroimage.usc.edu/brainstorm



(a) $\lambda = 0$

(b) Isotropic, $\lambda = 100$

(c) Anisotropic $\lambda = 500$

(d) Anisotropic $\lambda = 2000$

Fig. 4.1: Examples of isotropic and anisotropic TV regularization for the computation of barycenters between four input densities. The weights $(w_k)_{k=1}^N$ are bilinear interpolation weights, so that it is for instance $w = (1, 0, 0, 0)$ on the top left corner and $(0, 0, 0, 1)$ on the bottom right corner.

We used a squared Euclidean metric (4.9) on the flattened hemisphere. Since the data is defined on an irregular graph, instead of (4.8), we use a graph-based discrete gradient. We denote $((i,j))_{(i,j)\in\mathcal{G}}$ the graph which connects neighboring electrodes. The gradient operator on the graph is

$$\forall\, p \in \mathbb{R}^n, \quad \mathcal{A}p \overset{\text{def.}}{=} (p_i - p_j)_{(i,j)\in\mathcal{G}} \in \mathbb{R}^{|\mathcal{G}|}.$$

The total variation on this graph is then obtained by using $J = \lambda \|\cdot\|_1$, the ℓ^1 norm, i.e. we use $\beta = 1$ in (4.8).

Figure 4.3 compares the naive ℓ^2 barycenters (i.e. the usual mean), barycenters obtained without regularization (i.e. $\lambda = 0$) and barycenters computed with an increasing regularization strength λ. The input histograms $(p_k)_k$ being very noisy, the use of regularization is important to make the area of significant activity emerge

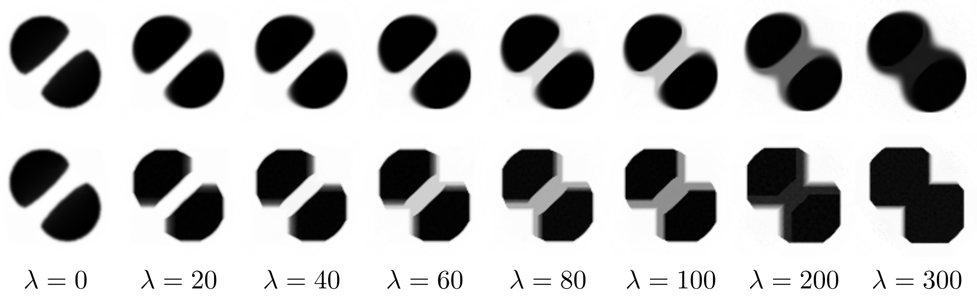


Fig. 4.2: Influence of λ parameter for the iso-barycenter (i.e. $w = (1/2, 1/2)$) between two input densities (they are the upper-left and upper-right corner of the $\lambda = 0$ case in Figure 4.1). *Top row:* isotropic total variation ($\beta = 2$). *Bottom row:* anisotropic total variation ($\beta = 1$).

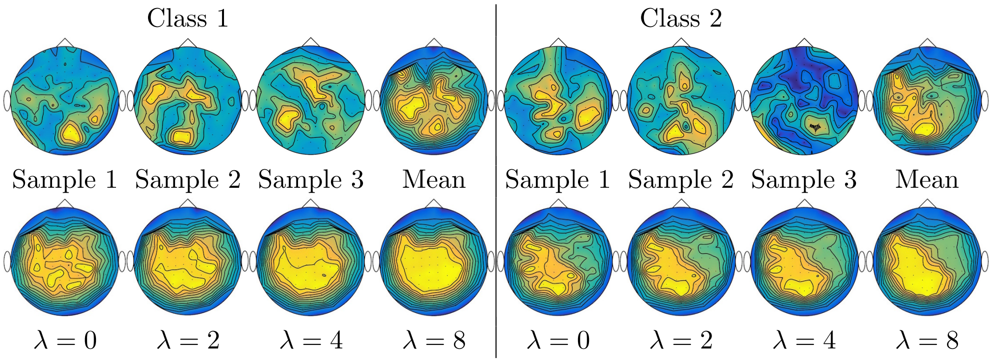


Fig. 4.3: Barycenter computation on MEG data. The left/right panels shows respectively the first and the second class, corresponding to recordings where the subject is asked to push the left or the right button. *Top row:* examples of input histograms q_k for each class, as well as the ℓ^2 mean $N^{-1} \sum_k q_k$. *Bottom row:* computed TV-regularized barycenter for different values of λ ($\lambda = 0$ corresponding to no regularization).

from the noise. The use of a TV regularization helps to keep a sharp transition between active and non-active regions.

4.6. Gradient Flow. Instead of computing barycenters, we now use our regularization to define time-evolutions, which are defined through a so-called discrete gradient flow.

Starting from an initial histogram $p_0 \in \Sigma_n$, we define iteratively

$$p_{k+1} \overset{\text{def.}}{=} \underset{p \in \Sigma_N}{\operatorname{argmin}} \; H_{p_k}(p) + \tau f(p). \tag{4.10}$$

This means that one seeks a new iterate at (discrete time) $k+1$ that is both close (according to the Wasserstein distance) to p_k and minimizes the functional f. In the following, we consider the gradient flow of regularization functionals as considered before, i.e. that are of the form $\tau f = J \circ \mathcal{A}$. Problem (4.10) is thus a special case of (4.1) with $N = 1$.

Letting $k \to +\infty$, one can informally think of p_k as a discretization of a time evolution evaluated at time $t = k\tau$. This method is a general scheme presented in much detail in the monograph [3]. The use of an implicit time-stepping (4.10) allows one to define time evolutions to minimize functionals that are not necessarily smooth, and this is exactly the case of the total variation semi-norm (since J is not differentiable). The use of gradient flows in the context of the Wasserstein fidelity to the previous iterate has been introduced initially in the seminal paper [18]. When f is the entropy functional, this paper proves that the countinous flow defined by the limit $k \to +\infty$ and $\tau \to 0$ is a heat equation. Numerous theoretical papers have shown how to recover many existing non-linear PDE's by considering the appropriate functional f, see for instance [25, 17].

The numerical method we consider in this article is the one introduced in [27], that makes use of the entropic smoothing of the Wasserstein distance. It is not the scope of the present paper to discuss the problem of approximating gradient flows and the underlying limit non-linear PDE's, and we refer to [27] for an overview of the vast literature on this topic. A major bottleneck of the algorithm developed in [27] is that it uses a primal optimization scheme (Dykstra's algorithm) that necessitates the computation of the proximal operator of f according to the Kulback-Leibler divergence. Only relatively simple functionals (basically separable functionals such as the entropy) can thus be treated by this approach. In contrast, our dual method can cope with a much larger set of functions, and in particular those of the form $f = J \circ \mathcal{A}$, i.e. obtained by pre-composition with a linear operator.

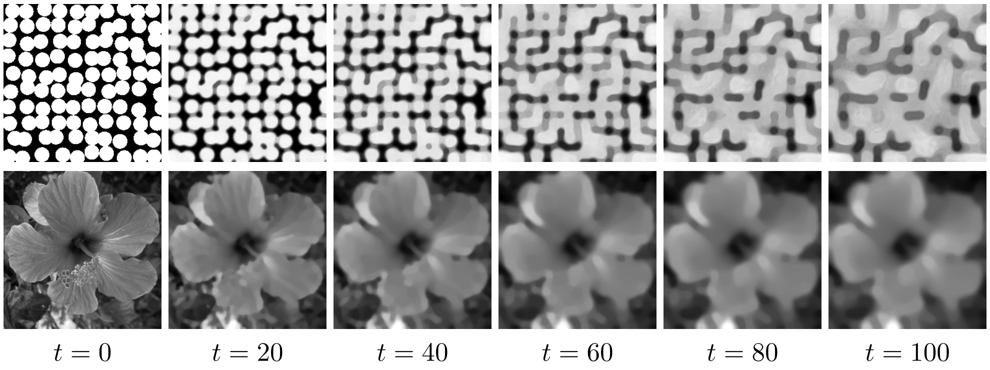


Fig. 4.4: Examples of gradient flows (4.10) at various times $t \overset{\text{def.}}{=} k\tau$.

Figure 4.4 shows examples of gradient flows computed for the isotropic total variation $f(p) = \|\nabla p\|_1$ as defined in (4.8). We use the discretization setup considered in Section 4.4. This is exactly the regularization flow considered by [12]. This paper defines formally the highly non-linear fourth order PDE corresponding to the limit flow. This is however not a "true" PDE since the initial TV functional is non-smooth, and derivatives should be understood in a weak sense, as limit of an implicit discrete time stepping. While the algorithm proposed in [12] uses the usual (unregularized) Wasserstein distance, the use of a regularized transport allows us to deal with problems of larger sizes, with a faster numerical scheme. The price to pay is an additional blurring introduced by the entropic smoothing, but this is acceptable for applications to denoising in imaging. Figure 4.4 illustrates the behavior of this TV regularization flow, which has the tendency to group together clusters of mass, and performs some kind of progressive "percolation" over the whole image.

Conclusion. In this paper, we introduced a dual framework for the resolution of certain variational problems involving Wasserstein distances. The key contribution is that the dual functional is smooth and that its gradient can be computed in closed form and involves only multiplications with a Gibbs kernel. We illustrate this approach with applications to several problems revolving around the idea of Wasserstein barycenters. This method is particularly advantageous for the computation of regularized barycenters, since pre-composition by linear operator (such as discrete gradient on images or graphs) of functionals is simple to handle. Our numerical findings is that entropic smoothing is crucial to stabilize the computation of barycenters and to obtain fast numerical schemes. Further regularization using for instance a total variation is also beneficial, and can be used in the framework of gradient flows.

Acknowledgments. The work of Gabriel Peyré has been supported by the European Research Council (ERC project SIGMA-Vision). Marco Cuturi gratefully acknowledges the support of JSPS young researcher A grant 26700002. We would like to thank Antoine Rolet, Nicolas Papadakis and Julien Rabin for stimulating discussions. We would like to thank Valentina Borghesani, Manuela Piazza et Marco Buiatti for giving us access to the MEG data. We would like to thank Fabian Pedregosa and the chaire "Économie des Nouvelles Données" for the help in the preparation of the MEG data.

Appendix A. Legendre Transform with Respect to Two Histograms.

Theorem 2.4 can be extended to study the Legendre transform of $W_\gamma(p,q)$ with respect to both arguments (p,q) instead of only p. Indeed, expression (2.5) shows that $(p,q) \mapsto W_\gamma(p,q)$ is a convex function (as a maximum of linear forms), so that one can define $\forall\,(g,h) \in \mathbb{R}^n \times \mathbb{R}^n$,

$$W_\gamma^*(g,h) = \max_{p,q\in\Sigma_n} \langle g,\, p\rangle + \langle h,\, q\rangle - W(p,q).$$

The following proposition adapts to this setting.

Proposition A.1. *The function W_γ^* is C^∞ at $(g,h) \in \mathbb{R}^n \times \mathbb{R}^n$ and, writing $K = e^{-M/\gamma}$, $\alpha = e^{g/\gamma}, \beta = e^{h/\gamma}$ and $\mathcal{K}_{\alpha\beta} = \operatorname{diag}(\alpha)K\beta$, we have that*

$$\begin{aligned} W_\gamma^*(g,h) &= -\gamma \log \alpha^T K\beta, \\ \nabla W_\gamma^*(g,h) &= \frac{1}{\alpha^T K\beta}\begin{bmatrix}\mathcal{K}_{\alpha,\beta}\\ \mathcal{K}_{\beta,\alpha}\end{bmatrix}, \\ \nabla^2 W_\gamma^*(g) &= \frac{1}{\gamma\alpha^T K\beta}\begin{bmatrix}A_\gamma(g,h) & B_\gamma(g,h)\\ B_\gamma(h,g) & A_\gamma(h,g)\end{bmatrix}. \end{aligned}$$

$$\textit{where}\quad \begin{cases} A_\gamma(g,h) & = \operatorname{diag}(\mathcal{K}_{\alpha\beta}) - \frac{1}{\alpha^T K\beta}\mathcal{K}_{\alpha\beta}\mathcal{K}_{\alpha\beta}^T, \\ B_\gamma(g,h) & = \operatorname{diag}(\beta)K\operatorname{diag}(\alpha) - \frac{1}{\alpha^T K\beta}\mathcal{K}_{\beta\alpha}\mathcal{K}_{\alpha\beta}^T. \end{cases}$$

Moreover, the gradient function $(g,h) \mapsto \nabla W_\gamma^*(g,h)$ is $2/\gamma$ Lipschitz.

Proof. One has that $W^*_\gamma(g,h)$ can be written

$$\begin{aligned}
&\max_{p,q\in\Sigma_n} \langle g,\, p\rangle + \langle h,\, q\rangle - \max_{u,v} \langle u,\, p\rangle + \langle v,\, q\rangle - \beta_{\gamma,M}(u,v)\\
&= \max_{p,q} \; - \max_{u,v} \langle u+g,\, p\rangle + \langle v+h,\, q\rangle - \beta_{\gamma,M}(u,v)\\
&= \max_{p,q} \; - \max_{u',v'} \langle u',\, p\rangle + \langle v',\, q\rangle - \beta_{\gamma,M}(u'+g, v'+h)\\
&= \max_{p,q} \; - W_{M+g\mathbb{1}^T+\mathbb{1}h^T}(p,q)\\
&= \max_{p,q} \; - \min_{X\in U(p,q)} \langle X,\, M-g\mathbb{1}^T - \mathbb{1}h^T\rangle - \gamma E(X)\\
&= - \min_{X\in\Sigma_{n^2}} \langle X,\, M-g\mathbb{1}^T - \mathbb{1}h^T\rangle - \gamma E(X).
\end{aligned}$$

One verifies that the last Eq. is equivalent to a classic maximal entropy problem which can be solved uniquely with a Gibbs distribution equal to $X^\star$ given below,

$$X^\star = \frac{\operatorname{diag}(\alpha)K\operatorname{diag}(\beta)}{\alpha^T K\beta}.$$

Substituting this expression in the formula above for $W^*_\gamma(g,h)$ yields that

$$W^*_\gamma(g,h) = -\gamma\log\alpha^T K\beta.$$

Since the gradients with respect to g and h of $W^*_\gamma(g,h)$ are $X^\star\mathbb{1}$ and $X^{\star T}\mathbb{1}$ respectively, this results in the expression provided above. The Hessian follows from that result, and the Lipschitz continuity of the gradient can be obtained by showing that the Hessian's trace can be upper-bounded by $2/\gamma$ by noticing that the trace of both $A_\gamma(g,h)$ and $A_\gamma(h,g)$ is upper-bounded by $\alpha^T K\beta$. □

REFERENCES

[1] I. Abraham, R. Abraham, M. Bergounioux, and G. Carlier. Tomographic reconstruction from a few views: a multi-marginal optimal transport approach. *Preprint Hal-01065981*, 2014.

[2] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM J. on Mathematical Analysis*, 43(2):904–924, 2011.

[3] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures.* Springer, 2006.

[4] A. Banerjee, S. Merugu, I. S Dhillon, and J. Ghosh. Clustering with Bregman divergences. *The Journal of Machine Learning Research*, 6:1705–1749, 2005.

[5] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces.* Springer-Verlag, New York, 2011.

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

[7] J-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[8] J. Bigot and T. Klein. Consistent estimation of a population barycenter in the Wasserstein space. *Preprint arXiv:1212.2562*, 2012.

[9] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

[10] N. Bonneel, M. van de Panne, S. Paris, and W. Heidrich. Displacement interpolation using lagrangian mass transport. *ACM Transactions on Graphics (SIGGRAPH ASIA'11)*, 30(6), 2011.

[11] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004.

[12] M. Burger, M. Franeka, and C-B. Schonlieb. Regularised regression and density estimation based on optimal transport. *Appl. Math. Res. Express*, 2:209–253, 2012.

[13] G. Carlier, A. Oberman, and E. Oudet. Numerical methods for matching for teams and Wasserstein barycenters. Preprint hal-00987292, Preprint HAL-00987292, 2014.

[14] V. Caselles, A. Chambolle, and M. Novaga. The discontinuity set of solutions of the TV denoising problem and some extensions. *SIAM Multiscale Modeling and Simulation*, 6(3):879–894, 2007.

[15] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300, 2013.

[16] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 685–693, 2014.

[17] U. Gianazza, G. Savaré, and G. Toscani. The wasserstein gradient flow of the Fisher information and the quantum drift-diffusion equation. *Archive for Rational Mechanics and Analysis*, 194(1):133–220, 2009.

[18] R. Jordan, D. Kinderlehrer, and O. Otto. The variational formulation of the Fokker-Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[19] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[20] J. Lellmann, D. Lorenz, C.-B. Schonlieb, and T. Valkonen. Imaging with kantorovich-rubinstein discrepancy. *to appear in SIAM Journal on Imaging Sciences*, 2015.

[21] B. Levy. A numerical algorithm for L^2 semi-discrete optimal transport in 3d. *M2AN, to appear*, 2015.

[22] J. Maas, M. Rumpf, C. Schonlieb, and S. Simon. A generalized model for optimal transport of images including dissipation and density modulation. *Arxiv preprint*, 2014.

[23] Q. Mérigot. A multiscale approach to optimal transport. *Computer Graphics Forum*, 30(5):1583–1592, 2011.

[24] F. Nielsen. Jeffreys centroids: A closed-form expression for positive histograms and a guaranteed tight approximation for frequency histograms. *IEEE Signal Processing Letters (SPL)*, 20(7), 2013.

[25] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in partial differential equations*, 26(1-2):101–174, 2001.

[26] O. Pele and M. Werman. Fast and robust earth mover's distances. In *ICCV'09*, 2009.

[27] G. Peyré. Entropic wasserstein gradient flows. Preprint 1502.06216, arXiv, 2015.

[28] J. Rabin and N. Papadakis. Convex color image segmentation with optimal transport distances. In *Proc. SSVM'15*, 2015.

[29] Y. Rubner, C. Tomasi, and L.J. Guibas. The earth mover's distance as a metric for image retrieval. *IJCV: International Journal of Computer Vision*, 40, 2000.

[30] B. Schmitzer and C. Schnorr. Object segmentation by shape matching with wasserstein modes. In *Proc. EMMCVPR'13*, volume 8081 of *Lecture Notes in Computer Science*, pages 123–136. Springer Berlin Heidelberg, 2013.

[31] J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (Proc. SIGGRAPH 2015)*, to appear, 2015.

[32] J. Solomon, R.M. Rustamov, L. Guibas, and A. Butscher. Wasserstein propagation for semi-supervised learning. In *Proc. ICML 2014*, 2014.

[33] P. Swoboda and C. Schnorr. Convex variational image restoration with histogram. *SIAM J. Imag. Sci.*, 6(3):1719?–1735, 2013.

[34] C. Villani. *Optimal transport: old and new*, volume 338. Springer Verlag, 2009.

[35] G-S. Xia, S. Ferradans, G. Peyré, and J-F. Aujol. Synthesizing and mixing stationary gaussian texture models. *SIAM Journal on Imaging Sciences*, 7(1):476–508, 2014.

[36] G. Zen, E. Ricci, and N. Sebe. Simultaneous ground metric learning and matrix factorization with earth movers distance. *Proc. ICPR'14*, 2014.