A Smoothed Dual Approach for Variational Wasserstein Problems

Marco Cuturi  
Graduate School of Informatics, Kyoto University

Gabriel Peyré  
CNRS and CEREMADE, Université Paris-Dauphine,

Antoine Rolet  
Graduate School of Informatics, Kyoto University

Abstract
Variational problems that involve Wasserstein distances have been recently proposed as a mean 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 Carlier et al. (2014) can be regularized using an entropic smoothing, which leads to smooth, differentiable, convex optimization problems that are simpler to implement and numerically more stable. In addition to such favorable properties, we propose a simple and effective heuristic to initialize variables with that formulation. We illustrate the versatility of our smoothed dual formulation by applying it to the computation of Wasserstein barycenters and by carrying out dictionary learning on a dataset of histograms using the Wasserstein distance as the fitting error.

1. Introduction
Histograms are popularly used in machine learning to represent complex objects as frequency vectors in the probability simplex. By defining first a set of relevant features, one can then form for each object a normalized histogram that keeps track of the frequencies of each of these features—e.g. bags-of-words for text (Salton et al., 1975), bags-of-visual-words for images (Lowe, 1999; Oliva & Torralba, 2001). To compare two histograms, information-based quantities such as the Hellinger and $\chi^2$ distances, the Kullback-Leibler and Jensen-Shannon divergences, have the advantages of being simple and fast to compute. Optimal transport distances (Villani, 2009, §7)—a.k.a Wasserstein or earth mower’s distances (Rubner et al., 2000)—require more computational effort but are more versatile: by incorporating in their definition a metric between the features themselves, they can compare sparse histograms even if their support do not overlap significantly, which can be crucial when the dictionary size is large. Their excellent performance (Pele & Werma, 2009, and references therein) comes, however, at a price: computing optimal transport distances requires solving a costly network flow problem, whose cost scales super-cubically with the dimension of the considered histograms. That cost becomes even more of a drawback if one attempts to learn from histogram data using the optimal transport geometry.

Variational Wasserstein problems  
Indeed, while many learning problems on histograms—such dictionary learning or clustering—can be framed as variational problems that involve distances between pairs of histograms, and have been successfully handled as such using Bregman divergences (Lee & Seung, 1999; Banerjee et al., 2005; Nielsen, 2013), solving such problems with the Wasserstein metric can be extremely challenging. Agueh & Carlier (2011) 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, Solomon et al. (2014) proposed the Wasserstein propagation-on-graphs framework, and showed that it involves a linear program of gigantic size, which is only feasibly solved in restrictive cases. 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 (Cuturi & Doucet, 2014), they have found usage in statistics to develop population estimators (Bigot & Klein, 2012) and in computer graphics to perform image modification (Xia et al., 2014; Bonneel et al., 2013). There is thus a need to provide fast and robust optimization techniques to solve Wasserstein variational problems and open new methodological perspectives. This is our aim in this paper, and we propose a general framework that exploits regularization, Legendre duality and the classical toolbox of convex optimization.

Related Approaches  
Rabin et al. (2012); Bonneel et al. (2013) proposed rough approximations of the Wasserstein distance to solve Wasserstein variational problems. Both can be used in practice to solve problems that involve clouds of points in low dimensions. Cuturi & Doucet (2014) propose to leverage the en-
tropic regularization of Wasserstein distances introduced by Cuturi (2013) to study Wasserstein variational problems that have differentiable objective functions. 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, Carlier et al. (2014) 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.

Contributions Our main contribution is to combine the strengths of the dual formulation of Carlier et al. with the smoothing strategy laid out by Cuturi & Doucet to obtain a smooth optimization problem whose objectives and derivatives can be computed in closed form in §2. We then proceed by applying this approach to compute Wasserstein barycenters in §3. We not only provide numerical evidence that our approaches make sense, we also show that the true minimum of discrete Wasserstein problems may be in fact extremely noisy. Therefore, our regularization may not only be favorable for computational reasons, it may also have beneficial effects in terms of modeling and regularity of the obtained solutions. Finally, we learn dictionaries from histogram data with the Wasserstein metric as a fitting criterion in §4, a task similar to that of topic modeling.

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, $d(u)$ is the diagonal matrix with diagonal $u$. $1_n \in \mathbb{R}^n$ is the (column) vector of ones.

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 $(X, D)$ such that $M_{ij} = D(x_i, x_j)\rho, \rho \geq 1$.

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

$$U(p, q) \equiv \{ X \in \mathbb{R}^n_+ ; X1_n = p, X^T1_n = q \},$$

$$E(X) \equiv -\sum_{ij} h(X_{ij}),$$

where $\forall x > 0, h(x) \equiv x \log x, h(0) = 0$. We follow Cuturi’s approach (2013) and introduce an entropy-regularized optimal transport problem:

$$W_\gamma(p, q) \equiv \min_{X \in U(p, q)} \langle M, X \rangle - \gamma E(X),$$

where $\gamma \geq 0$, and For $\gamma = 0$, one recovers the usual optimal transport problem, which is a linear program. $W_0$ is known as the Wasserstein (or EMD) distance between $p$ and $q$. For $\gamma > 0$, Problem (2) is strongly convex and hence admits a unique optimal coupling $X^\gamma_\ast$. Cuturi (2013) called the resulting cost $\langle M, X^\gamma_\ast \rangle$ the Sinkhorn divergence between $p$ and $q$. While $X^\gamma_\ast$ is not necessarily unique for $\gamma = 0$, we show in the following proposition that in the small $\gamma$ limit, the regularization captures the maximally entropic coupling.

Proposition 1. One has $W_\gamma \to W_0$ as $\gamma \to 0$, and denoting $X^\gamma_\ast$ the unique solution of (2), one has

$$X^\gamma_\ast \to X^0_\ast = \arg\max_X \{ E(X) ; \langle M, X \rangle = W_0(p, q) \} .$$

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

$$0 \leq \langle M, X_\ell \rangle - \langle M, X \rangle \leq \gamma_\ell (E(X_\ell) - E(X)) .$$

Since $E$ is continuous, taking the limit $\ell \to +\infty$ in this expression shows that $\langle M, X^\ast \rangle = \langle M, X \rangle$ so that $X^\ast$ is a feasible point of (3). Furthermore, dividing by $\gamma_\ell$ in (4) and taking the limit shows that $E(X) \leq E(X^\ast)$, which shows that $X^\ast$ is a solution of the maximization (3). Since the solution $X^0_\ast$ to this program is unique by strict convexity of $-E$, one has $X^\ast = X^0_\ast$, and the whole sequence is converging. □
Cuturi & Doucet (2014) provided a dual expression for \( W_\gamma \). The proof of that result follows from an application of Fenchel-Rockafellar duality to the primal problem (2). The indicator function of a closed convex set \( C \) is \( \iota_C(x) = 0 \) for \( x \in C \) and \( \iota_C(x) = +\infty \) otherwise.

**Proposition 2.** One has
\[
W_\gamma(p,q) = \max_{u,v \in \mathbb{R}^n} \langle u, p \rangle + \langle v, q \rangle - B(u,v),
\]
where
\[
B(u,v) = \begin{cases} \gamma \sum_{i,j} \exp\left(\frac{1}{\gamma}(u_i + v_j - M_{ij})\right), & \text{if } \gamma > 0; \\ \iota_C(u,v), & \text{if } \gamma = 0, \end{cases}
\]
where
\[
C_M = \{(u,v) : u_i + v_j \leq M_{ij}\}.
\]

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 \( \gamma \geq 0 \). When \( \gamma > 0 \), \( H_q \) has the following properties, which follow from the direct differentiation of expression (5):

**Proposition 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 \( \nabla H_q(p) = u^* \) where \( u^* \) is the unique solution of (5) satisfying \( \langle u^*, 1_n \rangle = 0 \).

Computing both \( H_q \) and its gradient requires thus the resolution of the optimization problem in Equation (5), which can be solved with a Sinkhorn fixed-point iteration (Cuturi, 2013) as remarked by Cuturi & Doucet (2014, §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. Carlier et al. (2014, 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.

We adapt first the result of Carlier et al. to our notations. Given a cost matrix \( M \in \mathbb{R}^{n \times n} \) and a vector \( g \in \mathbb{R}^n \), we introduce for \( i \leq n \) the set \( N_{M,g}(i) = \arg\min_k M_{ik} - g_i \). 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 \( \delta_j \) (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 Carlier et al.:

**Proposition 4** (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 \( g \in \mathbb{R}^n \),
\[
S_q(g) \overset{\text{def}}{=} \sum_{i \leq n} q_i \sigma_{M,g}(i) \in \partial H_q^*(g).
\]

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

The main result of this paper is to show that the Legendre transform of the smoothed Wasserstein distance \( \gamma > 0 \) has a closed form. 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 costly optimization problem. These properties are at the core of the computational framework we develop in this paper.

**Theorem 1** (Legendre Transform of \( H_q \)). For \( \gamma > 0 \), the Fenchel-Legendre dual function \( H_q^* \) is \( C^\infty \). 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 \( K = e^{-M/\gamma} \),
\[
H_q^*(g) = \gamma \langle E(q) + \langle g, \log K \alpha \rangle \rangle, \quad \nabla H_q^*(g) = \alpha \circ \left( K \frac{q}{K \alpha} \right) \in \Sigma_n, \quad \nabla^2 H_q^*(g) = \frac{1}{\gamma} d \left( \alpha \circ K \frac{q}{K \alpha} \right) - \frac{1}{\gamma} d(\alpha) K d \left( \frac{q}{(K \alpha)^2} \right) K d(\alpha).
\]
\[ H^*_q,M(g) = \max_{p \in \Sigma_n} \langle g,p \rangle - \max_{u,v} \langle u,p \rangle + \langle v,q \rangle - B(u,v) \]
\[ = \max_{p \in \Sigma_n} - \max_{u',v} \langle u',p \rangle + \langle v,q \rangle - B(u' + g, v) \]
\[ = \max_{p \in \Sigma_n} - H_{q,M} - \gamma \tau(p) \]
\[ = -\min_{p \in \Sigma_n} \min_{X \in U(p,q)} \langle M - g1^T, X \rangle - \gamma E(X). \]

This leads to an optimal transport problem which is only constrained by one marginal,
\[ H^*_q,M(g) = \min_{\lambda} \langle M - g1^T, X \rangle - \gamma E(X) \]
which can be explicitly solved by writing first order conditions for (9) to obtain that, at the optimum, we necessarily have \( \log(X^*_j) = \frac{1}{\gamma}(g_j - M_{ij} - \rho_j) - 1 \) for some vector of values \( \rho \in \mathbb{R}^d \). Therefore \( X^* \) has the form \( X^* = d(\alpha)K d(q/K\alpha) \), using the notation \( \alpha = e^{q/K}\gamma \). Because of the marginal constraint that \( X^*1 = q \), the rightmost diagonal matrix must necessarily be equal to \( d(q/K\alpha) \), and thus \( X^* = d(\alpha)K d(q/K\alpha) \). Therefore, the Legendre transform \( H^*_q,M \) has a closed form, \[ H^*_q,M(g) = -\langle M - g1^T, X^* \rangle + \gamma E(X^*) \]
which can be simplified to
\[ H^*_q,M(g) = -\gamma 1^T \left( (K\alpha) \circ h(q/K\alpha) \right) \]
by using the fact that \( X^* = d(\alpha)K d(q/K\alpha) \). This equation can be simplified further to obtain the expression provided in Equation (6). Using Equation (10), we have that
\[ \nabla H^*_q,M(g) = X^*1 = \alpha \circ \left( K \frac{q}{K\alpha} \right). \]

Computations for the Hessian follow directly from this expression and result in Equation (8). 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 \( \nabla H^*_q,g \) is in the simplex), which proves the \( \frac{1}{\gamma} \)-Lipschitz continuity of the gradient of \( H^*_q \).

**Remark 1.** In some settings, such as the Wasserstein propagation framework of Solomon et al. (2014), the Wasserstein distance with respect to two arguments might have to be minimized. We provide the formulation for the corresponding Legendre transform in Theorem 6 in the Appendix.

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 Agueh & Carlier (2011), Cuturi & Doucet (2014) introduced the smoothed WBP with \( \gamma \)-entropic regularization (\( \gamma \)-sWBP) as
\[ \min_{p \in \Sigma_n} \sum_{k=1}^N \lambda_k H_{q,k}(p) \]  
where \((q_1, \ldots, q_N) \) is a family of histograms in \( \Sigma_n \). When \( \gamma = 0 \), the \( \gamma \)-sWBP is exactly the WBP. When \( \gamma > 0 \) the \( \gamma \)-sWBP is a strictly convex optimization problem that admits a unique solution. Cuturi & Doucet proposed to carry out a simple a gradient descent approach to solve it. They show that the \( N \) gradients \( \{\nabla H_{q,k}(p)\}_{k=1,N} \) can be computed by solving \( N \) Sinkhorn fixed point iterations. 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 their approach. We take here a different route to solve the \( \gamma \)-sWBP, which can be either interpreted as a smooth alternative to the dual WBP studied by Carlier et al. (2014), or the dual counterpart to the smoothed WBP of Cuturi & Doucet.

**Theorem 2.** The barycenter \( p^* \) solving (11) satisfies
\[ \forall k = 1, \ldots, N, \quad p^* = \nabla H^*_{q,k}(g_k) \]
where \((g_k)_k \) are any solution of the smoothed dual WBP:
\[ \min_{g_1, \ldots, g_N} \sum_k \lambda_k H_{q,k}(g_k) \quad \text{s.t.} \quad \sum_k \lambda_k g_k = 0. \]  

Proof. We re-write the barycenter problem
\[ \min_{p_1, \ldots, p_N} \sum_k \lambda_k H_{q,k}(p_k) \quad \text{s.t.} \quad p_1 = \ldots = p_N \]
whose Fenchel-Rockafellar dual reads
\[ \min_{g_1, \ldots, g_N} \sum_k \lambda_k H^*_{q,k}(g_k/\lambda_k) \quad \text{s.t.} \quad \sum_k g_k = 0. \]
Since the primal problem is strictly convex, the primal-dual relationships show that the unique solution \( p^* \) of
the primal can be obtained from any solution \((g^*_k)_k\) via the relation
\[ p^*_k = \nabla H^*_{q_k^*}(\tilde{g}^*_k/\lambda_k). \]
One obtains the desired formulation using the change of variable
\[ g_k = \tilde{g}^*_k/\lambda_k. \]

Theorem 2 provides a simple approach to solve the \(\gamma\)-sWBP: rather than minimizing directly the sum of regularized Wasserstein distances in Equation (11), this formulation only involves minimizing a strictly convex function with closed form objectives and gradients. The computation of its objective, gradient and Hessians with respect to all variables \(\{g_1, \cdots, g_N\}\) can be efficiently parallelized, as we show below.

### 3.2. Parallel Implementation

The objectives, gradients and Hessians of the Fenchel-Legendre dual \(H^*\) can be computed efficiently in the sense that they only involve 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 \((H^*_q(g_k))_k\), \(N\) gradients \((\nabla H^*_q(g_k))_k\) and \(N\) applications of the Hessian matrix to an arbitrary vector \(x_k\), \((\nabla^2 H^*_{q_k}(g_k)x_k)_k\), 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 have, using the following \(n \times N\) auxiliary matrices:

\[
A \triangleq e^{-G/\gamma}, \quad B \triangleq KA, \quad C \triangleq Q/B, \quad \Delta \triangleq A \circ (KC).
\]

\[
H^* \triangleq [H^*_q(g_1), \ldots, H^*_q(g_N)] = -\gamma 1^T_n \left( Q \circ \log(C) \right).
\]

\[
\nabla H^* \triangleq [\nabla H^*_q(g_1), \ldots, \nabla H^*_q(g_K)] = \Delta.
\]

\[
\nabla^2 H^*X = \frac{1}{\gamma} \left[ \Delta \circ X - A \circ \left( K \left( \frac{Q \circ (K(A \circ X))}{B \circ B} \right) \right) \right].
\]

### 3.3. Algorithm

The smoothed Dual WBP in Equation (13) has a smooth objective with respect to each of its variables \(g_k\), with a simple linear equality constraint. 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 descent, such as for instance those proposed by Nesterov (2007); Beck & Teboulle (2009), quasi-Newton or truncated Newton methods (Boyd & Vandenberghe, 2004, §10). In the latter case, the resulting KKT linear system is sparse, and solving it with pre-conjugate gradient techniques can be efficiently carried out. We omit details for lack of space and only report in this work fixed steplength first-order descent and results using L-BFGS. From the dual iterates \(q_k\) stored in a \(n \times N\) matrix \(G\), one recovers primal iterates using the formula (12), namely

\[ p_k = e^{q_k/\gamma} \circ K \frac{q_k}{K e^{q_k/\gamma}}. \]

At each intermediary iteration one can thus form a solution to the smoothed Wasserstein barycenter problem by averaging these primal solutions,

\[ \hat{p} = \Delta 1_N/N. \]

Upon convergence, these \(p_k\) are all equal to the unique solution \(p^*\). The average at each iteration \(\hat{p}\) converges towards that unique solution, and we use the sum of all line wise standard deviations of \(\Delta: 1^T_N \sqrt{(Z \circ Z)1_N/N}\), where \(Z = \Delta(I_N - \frac{1}{N} 1_N1_N^T)\) to monitor that convergence in our algorithms.

#### Algorithm 1 Smoothed Wasserstein Barycenter, Generic Algorithm

1. **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 \(\varepsilon > 0\).

2. **initialize** \(G \in \mathbb{R}^n \times N\) and form the \(n \times n\) matrix \(K = e^{-M/\gamma}\).

3. **repeat**

4. From gradient matrix \(\Delta\) (see Equation 14) produce update matrix \(\Delta\) using either \(\Delta\) directly or other methods such as L-BFGS.

5. \(G = G - \tau \Delta, \) update with fixed step length \(\tau\) or approximate line search to set \(\tau\)

6. \(G = G - \frac{1}{\|\lambda\|^2}(G\lambda)\lambda^T\) (projection such that \(G\lambda = 0\))

7. **until** \(1^T_N \sqrt{(Z \circ Z)1_N/N} < \varepsilon\), where \(Z = \Delta(I_N - \frac{1}{N} 1_N1_N^T)\)

8. **output** barycenter \(p = \Delta 1_N/N\).

### 3.4. Initialization Heuristic

Definition 1 provides an initialization heuristic to initialize both the primal and dual smoothed WBP, motivated by the fact that they provides directly the optimal primal and dual solutions when the histograms are Dirac histograms as proved in Proposition 5

**Definition 1** (Primal and Dual WBP Initialization). Let \((q_1, \cdots, q_N)\) be \(N\) target histograms in the simplex \(\Sigma_n\) and \(\lambda\) a vector of weights in \(\Sigma_N\). Let \(\bar{q} = \sum_k \lambda_k q_k \in \Sigma_n\). Define \(\kappa_\gamma\) as

\[
\kappa_\gamma = \begin{cases} 
e^{-M\bar{q}/\gamma}/(1_n^T e^{-M\bar{q}/\gamma}) & \text{if } \gamma > 0, \\ 
\delta_j, \text{ where } j \in \arg\min_k [M\bar{q}]_k, & \text{if } \gamma = 0. 
\end{cases}
\]
For $\gamma \geq 0$, the $\gamma$-smoothed primal and dual WBP can be initialized respectively with the following primal and $N$ dual feasible solutions:

$$p^{(0)} \overset{\mathrm{def}}{=} \kappa \gamma,$$

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

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

**Remark 2.** Notice that the primal initialization we propose is different depending on the value of $\gamma$. For $\gamma > 0$, $\kappa$ is the normalized vector of row-wise geometric averages, with weights $\bar{q}$, of the columns of the matrix $K = e^{-M/\gamma}$ whereas for $\gamma = 0$, $\kappa$ is a vector of zero values except for a value of $K$ corresponding to the (or any, if many) smallest entry of $M\bar{q}$. On the other hand, our dual initialization is the same regardless of the fact that $\gamma$ is null or not, namely it applies to both smoothed and non-smoothed Wasserstein barycenter problems.

The initializations proposed in Definition 1 can be interpreted as simple solutions of the WBP in the degenerate case that all histograms are Dirac histograms:

**Proposition 5.** Let $(q_1, \cdots, q_N)$, be Dirac histograms, namely histograms that are zero everywhere but for one coordinate equal to 1, paired with a vector $\lambda$ of weights in $\Sigma_N$. For $\gamma \geq 0$, the $\gamma$-sWBP primal and dual problems are solved exactly using the initialization described in Definition (1).

**Proof.** To simplify notations, we write $p = p^{(0)}$ and $g_k = g_k^{(0)}$ as defined in Definition 1 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 $1 \leq k \leq N$. Writing $\alpha_k = e^{q_k/\gamma}$, we recover that

$$\alpha_k = \frac{\kappa}{\xi_k},$$

where $\xi_k \overset{\mathrm{def}}{=} e^{-M q_k/\gamma}$. 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, $\xi_k \overset{\mathrm{def}}{=} e^{-M q_k/\gamma} = K_j$, where $K_j$ is the $j^{th}$ column of the matrix $K = e^{-M/\gamma}$. Therefore,

$$\alpha_k = \frac{\kappa}{K_j}.$$

Let us now compute the gradient $\nabla_k$ of $H_{q_k}$ at $g_k$ by following Equation (7):

$$\nabla_k = \alpha_k \circ \left( K q_k / K \alpha_k \right).$$

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

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

Since only the $j^{th}$ element of $q_k$ is non-zero by definition,

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

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

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

which holds for all indices $1 \leq k \leq N$.

**3.5. Smoothing and Stabilization of the WBP**

Before showing the numerical advantages of our smoothing approach in §3.6, we discuss first the practical implications of solving the smoothed WBP rather than the original WBP. We make the perhaps surprising claim that smoothing the WBP is not only beneficial computationally, it also seems to provide better behaved barycenters.

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 $N$:

$$\min_{X_1, \cdots, X_N, p} \sum_{k=1}^N \lambda_k \langle X_k, M \rangle$$

s.t. $X_k \in \mathbb{R}^{n \times n}_+$, $\forall k \leq N, p \in \Sigma_n$, $X_k^T 1_n = q_k, \forall k \leq N$, $X_1 1_n = \cdots = X_N 1_n = p$. 

Given optimal couplings $X_1^*, \ldots, X_N^*$, the solution to the WBP is equal to the marginal common to all those couplings, namely $p = X_k^* 1_n$ for any $k \leq N$. 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 $(\mu_1, \sigma_1)$ and $(\mu_2, \sigma_2)$ respectively is a Gaussian of mean $(\mu_1 + \mu_2)/2$ and standard deviation $(\sigma_1 + \sigma_2)/2$ (Agueh & Carlier, 2011, §6.3). This fact is illustrated in the top-left plot of Figure 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 $p^*$, displayed in the bottom-left plot, is the exact optimal solution of Equation (18), computed with the simplex method. That WBP reduces to a linear program of $2 \times 100^2$ variables and 300 constraints. We observe that $W_2^2(p^*, q_1) + W_2^2(p^*, q_2) = 0.5839950$ 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 $p^*$ converges to the continuous barycenter as the grid size tends to zero, as shown in (Carlier et al., 2014). This observation illustrates however that, because it is defined as the argmin of a linear program, the true Wasserstein barycenter may be extremely unstable. Regularization the Wasserstein distances has thus the added benefit of smoothing the resulting solution, which may have in many applications a beneficial effect.

### 3.6. Experiments

We describe first the two direct competitors of our optimization framework, the smooth primal approach of Cuturi & Doucet and the dual approach of Carlier et al.. We compare them with our smooth dual approach to compute the Wasserstein barycenter of 12 histograms laid out on the $100 \times 100$ grid, as displayed in Figure 3.

**Smooth primal first-order descent** Cuturi & Doucet (2014, §5) proposed to minimize directly Equation (11) 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 $\alpha_k$ are the left scalings obtained with that subroutine. A weakness of that approach is that a precision threshold $\epsilon$ for the Sinkhorn fixed-point algorithm must be chosen. That precision can be measured by the difference in $l_1$ norm between the row and column marginals of $d(\alpha_k)e^{-M/\gamma} d(\beta_k)$ and targeted $p$ and $q_k$. Setting that tolerance $\epsilon$ to a large value ensures a faster convergence of the subroutine but noisy gradients and

![Figure 1](image1.png)

**Figure 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 18 (bottom right) barycenter computed with our smoothing approach.

![Figure 2](image2.png)

**Figure 2.** Plots of the exact barycenters for varying grid size $n$. 
Figure 5. Barycenters obtained for the three different techniques using the data described in Figure 3 after at most $10^4$ iteration units, each iteration unit being equal to $n^2N$ operations, here $(100 \times 100)^2 \times 12$.

Figure 3. 12 measures, truncated mixtures of Gaussians, used in our benchmark. Convergence speed results displayed in Figure 4 and barycenters obtained in 5.

therefore slower convergence of the WBP. Because the smoothed dual approach only relies on closed form expressions we do not have to take into account such a trade-off.

Smooth dual, first order & L-BFGS The dual formulation with variables $(q_1, \cdots, q_N) \in (\mathbb{R}^p)^N$ of Equation (13) with no regularization can be solved using At each iteration of that minimization, we can recover a feasible solution $p$ to the primal problem of Equation (11) via the primal-dual relation $p = \frac{1}{N} \sum_k \nabla H^*_q(g_k)$.

Dual descent with L-BFGS This approach amounts to solving directly the (non-differentiable) dual problem described in Equation (13) with no regularization, namely $\gamma = 0$. Subgradients for the Fenchel-Legendre transforms $H^*_q$ can be obtained in closed form through Proposition 4. As with the smoothed-dual formulation, we can also obtain a feasible primal solution by averaging subgradients.

We follow Carlier et al.’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 In all experiments below, we use as a unit of computation $Nh^2$ elementary operations. We plot the optimality gap (computed by considering the minimum value among all techniques obtained with $10^5$ iterations) w.r.t the optimum. Because the two smooth approaches ($\gamma > 0$) optimize a different objective than the one considered by the dual approach ($\gamma = 0$), we cannot compare smooth and non-smooth approaches directly. For the sake of completeness, we propose to show both smoothed ($\gamma = 1/100$) and non-smoothed objectives ($\gamma = 0$) for all techniques, keeping in mind that smoothed (primal and dual) techniques only optimize the former, whereas the dual descent optimizes the latter. Perhaps surprisingly, the smoothed dual approach presented here, and which aim at minimizing or $W_{1/100}$ distances, performs orders of magnitude better (both in speed and optimality gap) on the non-smoothed barycenter minimization task that only involves minimizing original $W_0$ distances. Note also that the initialization trick detailed in Definition 1 is effective for the dual formulation but does not seem to improve the primal formulation.

We also plot the solutions of all 3 algorithms in Figure 5 after up to $10^4$ iterations. Because the smoothed-primal and our smoothed-dual approach aim at minimizing the same objective, it is not surprising that their solutions are similar. Note however that with a budget of at most $10^4$ iterations the solution obtained with dual smoothing is more detailed than the one obtained with a primal descent. The right-most figure,
the smallest possible objective found after running 10 elements, namely find a dictionary histograms as a linear combination of the dictionary on a dataset of histograms arranged as a binary learning with a Wasserstein reconstruction error rem 2, and how it can be adapted to carry out dictionaries, as was also observed by the authors themselves.

The original Wasserstein barycenter problem formulation proposed in §3.4. Solutions displayed after convergence in Figure 5 obtained following Carlier et al.’s approach, shows that the original Wasserstein barycenter problem formulation, without smoothing, can yield very irregular solutions, as was also observed by the authors themselves in their paper.

4. Dictionary learning with smooth Wasserstein fitting cost

We illustrate in this section the versatility of Theorem 2, and how it can be adapted to carry out dictionary learning with a Wasserstein reconstruction error on a dataset of histograms arranged as a $n \times N$ matrix $Q = [q_1, \cdots, q_N]$.

Problem Formulation The dictionary learning problem with Wasserstein fitting cost consists in learning simultaneously $m$ dictionary elements and $N$ vectors of weights in $\Sigma_m$ to reconstruct each of the original histograms as a linear combination of the dictionary elements, namely find a dictionary $D$ and weights $\Lambda$ such that

$$\min_{D \in \mathbb{R}^{n \times m}, \Lambda \in \mathbb{R}^{m \times N}} \sum_{k=1}^{N} H_{q_k}(D\Lambda_k)$$

with the implicit constraint that $D\Lambda \in \Sigma_m^n$.

Previous work This problem was previously considered by Sandler & Lindenbaum (2009) and Zen et al. (2014). Sandler & Lindenbaum (2009) rely on a fast approximation of optimal transport that uses Wavelet decompositions (Shirdhonkar & Jacobs, 2008) but which only works in practice when $X$ is $\mathbb{R}^2$ or $\mathbb{R}^3$. Zen et al. (2014) propose a similar matrix factorization and approach while also learning a cost matrix $M$ (not necessarily a metric) using labels. Their formulation relies on large-scale linear programs (not network flows) whose complexity grows very quickly with the relevant dimensions $m, N$ and $n$ of the problem. Our approach is not restricted to particular choices for $X$ and can scale to large dimensions/datasets/dictionary sizes: we do make any assumption on the metric space $(X, D)$ other that the $n^2$ matrix of pairwise distances between the $n$ points in $X$ can be stored in memory. This versatility is illustrated in our experiments below, in which we carry out Wasserstein NMF on texts seen as large clouds of points in $\mathbb{R}^{50}$ using word embeddings (Pennington et al., 2014; Zou et al., 2013).

Dual formulation Let $D^*_\Lambda$ (resp. $\Lambda^*_D$) be the solution of (19) with $\Lambda$ (resp. $D$) fixed. The following theorem links $D^*_\Lambda$ (resp. $\Lambda^*_D$) and $\Lambda$ (resp. $D$) when the rank of $\Lambda$ (resp. $D$) is $m$.

**Theorem 3.** $\Lambda^*_D$ can be computed as $D^*_\Lambda = \nabla H_{q_k}(g_i^{\Lambda^*_D})$ where and $\Lambda^*_D = \nabla H_{q_k}(g_i^\Lambda)$ is the solution of

$$\min_{D \in \mathbb{R}^{n \times m}} H_{q_k}(g_k)$$

such that $k = 1 \cdots N$.

Furthermore $D^*_\Lambda$ can be computed as

$$D^*_\Lambda = \nabla H_{q_k}(g_i^{\nabla H_{q_k}(g_i^\Lambda)})_{k=1 \cdots N}$$

where $G^{\nabla D}$ is the solution of

$$\min_{G \in \mathbb{R}^{n \times m}} \sum_{k=1}^{N} H_{q_k}(g_k)$$

The proof is almost identical to that of 2 and left out for lack of space.

When $D$ is fixed, $\Lambda^*_D$ can be computed with Algorithm 1 by replacing the projection step by $G = G - GDD^+$ where $D^+$ is the pseudo-inverse of $D$, and the output is $\Lambda^*_D = D^+ \nabla H_{q_k}(G)$. Similarly when $\Lambda$ is fixed, $D^*_\Lambda$ can be computed with Algorithm 1 by replacing the projection by $G = G - \Lambda^+ \Lambda G$, and the output is $D^*_\Lambda = \nabla H_{q_k}(G)\Lambda^+$. 

\[Figure 4.\] 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 nonsmooth metrics. Note the crucial importance of the initialization proposed in §3.4. Solutions displayed after convergence in Figure 5

\[Dual formulation \] Let $D^*_\Lambda$ (resp. $\Lambda^*_D$) be the solution of (19) with $\Lambda$ (resp. $D$) fixed. The following theorem links $D^*_\Lambda$ (resp. $\Lambda^*_D$) and $\Lambda$ (resp. $D$) when the rank of $\Lambda$ (resp. $D$) is $m$.

**Theorem 3.** $\Lambda^*_D$ can be computed as $D^*_\Lambda = \nabla H_{q_k}(g_i^{\Lambda^*_D})$ where and $\Lambda^*_D = \nabla H_{q_k}(g_i^\Lambda)$ is the solution of

$$\min_{D \in \mathbb{R}^{n \times m}} H_{q_k}(g_k)$$

such that $k = 1 \cdots N$.

Furthermore $D^*_\Lambda$ can be computed as

$$D^*_\Lambda = \nabla H_{q_k}(g_i^{\nabla H_{q_k}(g_i^\Lambda)})_{k=1 \cdots N}$$

where $G^{\nabla D}$ is the solution of

$$\min_{G \in \mathbb{R}^{n \times m}} \sum_{k=1}^{N} H_{q_k}(g_k)$$

The proof is almost identical to that of 2 and left out for lack of space.

When $D$ is fixed, $\Lambda^*_D$ can be computed with Algorithm 1 by replacing the projection step by $G = G - GDD^+$ where $D^+$ is the pseudo-inverse of $D$, and the output is $\Lambda^*_D = D^+ \nabla H_{q_k}(G)$. Similarly when $\Lambda$ is fixed, $D^*_\Lambda$ can be computed with Algorithm 1 by replacing the projection by $G = G - \Lambda^+ \Lambda G$, and the output is $D^*_\Lambda = \nabla H_{q_k}(G)\Lambda^+$. 

\[Figure 4.\] 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 nonsmooth metrics. Note the crucial importance of the initialization proposed in §3.4. Solutions displayed after convergence in Figure 5
We can minimize Equation (19) using alternated optimization. Our algorithm gains numerical stability when the columns of $D$ and $\Lambda$ are projected on the simplex (of size $n$ and $m$) at each iteration.

Note that if $D^T \mathbf{1} = 1$ and the rank of $\Lambda$ (resp. $D^T$) stays equal to $m$, $D^T \mathbf{1} = 1$ (resp. $\Lambda^T \mathbf{1} = 1$).

**Experiments on synthetic data** In this experiment with toy-data, we consider an input matrix $Q$ using 3 Gaussians centered around $-6, 0$ and 6 and with variance 2, discretized on $[-12, 12]$ with 401 evenly spaced points. Each column of $Q$ is generated as a weighted sum of those gaussians with a shift on the means. The weights (resp. shifts) are evenly distributed on $[0, 1]$ (resp. $[-2, 2]$) and each resulting vector is normalized so that its coordinates sum to one (examples of such histograms are provided in the top right part of Figure 6). We generate $N = 1000$ histograms that way. The lower half of Figure 6 displays dictionary learned under the Wasserstein metric with our approach (smoothing parameter $\gamma = 1/50$) and the standard NMF algorithm with multiplicative updates and KL fitting cost (Lee & Seung, 1999). These computations require less than a few minutes on a desktop machine.

![Figure 6](image)

**Figure 6.** (top left) gaussians used to generate the data (top right) 4 examples of histograms generated using random shifts and weights. We use $N = 1000$ histograms to learn a dictionary using either: (bottom left) our algorithm with Wasserstein fitting cost with $m = 3$ dictionary elements and $\gamma = 1/50$; (bottom right) a KL fitting cost with $m = 3$.

**Translingual Topic Modeling** We consider the problem of summarizing a set of texts seen as bags-of-words in a given language as a convex combinations of topics (bags-of-words) in a different language. To carry out such a task with optimal transport, we need a metric between words in both languages. This metric can be defined using Zou et al.’s work, which proposes a bilingual embedding of both English and Chinese words in $\mathcal{X} = \mathbb{R}^{50}$. The Euclidean metric between such embeddings can be used directly to define the cost matrix $M$, with an exponent $\rho = 1$. In this setting $n$, the total number of points, is equal to $n_E + n_C$, where $n_E$ (resp. $n_C$) is the size of the English (resp. Chinese) dictionary considered for such texts.

We applied our Wasserstein-NMF method to a all 10788 English texts taken from the reuters corpus. We remove common english stop-words and words that are not in the word embedding dictionary of Zou et al. (2013) to obtain dictionary sizes of $n_E = 11978$ English words and $n_C = 4070$ Chinese words. We set $m = 6$, $\gamma = \frac{1}{50}$ and we learn dictionary elements with the constraint that they are exclusively supported on the $n_C$ bins that correspond to Chinese words, namely we wish to reconstruct each English text (seen as a bag-of-English-words) using convex combinations of bags-of-Chinese-words. We display the first factor computed by our method in Figure 7 as a cloud-of-words, where each word’s size is proportional to its frequency, and only the most frequent words are reported. That topic is dominated by words (single Chinese characters or pairs) related to international trade.

![Figure 7](image)

**Figure 7.** Wordle representation of 1 of the 6 mandarin topics learned for $\gamma = 1/50$. Only the 50 most frequent words are represented. The font size increases monotonically with the word frequency.

**Conclusion**

In this paper, we introduced a dual framework for the resolution of certain variational problems involving Wasserstein distances. We illustrate this approach with two important problems, one that involves computing Wasserstein barycenters and another that involves learning a dictionary and weights with a Wasserstein fit. Our approach has several attrac-
tive qualities: (i) our entropic regularization ensures the
unicity of the optimal solution in the simple WBP
problem and facilitates the computation of each of the
convex sub-problems considered in dictionary learning;
(ii) we observe that solutions obtained with this regu-
larization exhibit a level of smoothness which is com-
parable to that of the original measures. This property
can be desirable in some cases. (iii) our approach can
be initialized very efficiently thanks to a simple rule
that is optimal in the simplified case where all original
measures are dirac masses. (iv) using Fenchel duality,
we show that Wasserstein variational problems can be
carried out using closed form functions. We believe
this class of approaches can be extended to more gen-
tal tasks and can scale up to more demanding learning
problems.

A. Appendix: Legendre Transform
with Respect to Two Histograms

Theorem 1 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 (5) shows
that \((p, q) \rightarrow W_\gamma(p, q)\) 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 6.** The function \( W_\gamma^* \) is \( C^\infty \) 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 \( K_{\alpha\beta} = d(\alpha)K\beta \), we have that

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

where
\[
\begin{align*}
A_\gamma(g, h) &= d(K_{\alpha\beta}) - \frac{1}{\alpha^T K\beta} K_{\alpha\beta} K_{\beta\alpha}^T, \\
B_\gamma(g, h) &= d(\beta) K d(\alpha) - \frac{1}{\alpha^T K\beta} K_{\beta\alpha} K_{\alpha\beta}^T.
\end{align*}
\]

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

**Proof.** One has that \( W_\gamma^*(g, h) \) can be written
\[
\max_{p, q \in \Sigma_n} \langle g, p \rangle + \langle h, q \rangle - \max_{u, v} \langle u, p \rangle + \langle v, q \rangle - \beta_{\gamma, M}(u, v)
\]
\[
= \max_{p, q} - \max_{u, v} \langle u + g, p \rangle + \langle v + h, q \rangle - \beta_{\gamma, M}(u, v)
\]
\[
= \max_{p, q} - \max_{u', v'} \langle u', p \rangle + \langle v', q \rangle - \beta_{\gamma, M}(u' + g, v' + h)
\]
\[
= \max_{p, q} - W_{M + g1^T + h1^T}(p, q)
\]
\[
= \max_{p, q} - \min_{X \in U(p, q)} \langle X, M - g1^T - h1^T \rangle - \gamma E(X)
\]
\[
= - \min_{X \in \Sigma_{\alpha\beta}} \langle X, M - g1^T - h1^T \rangle - \gamma E(X).
\]

One verifies that the last equation is equivalent to a
classic maximal entropy problem which can be solved
uniquely with a Gibbs distribution equal to \( X^* \) given
below,
\[
X^* = \frac{d(\alpha)K d(\beta)}{\alpha^T K\beta}.
\]

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 \( X^*1 \) and \( X^*1 \) respectively, this results in the expres-
sion 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 \(1/\gamma\) 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**

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

Banerjee, Arindam, Merugu, Srujana, Dhillon, Inderjit
S. and Ghosh, Joydeep. Clustering with Bregman
divergences. *The Journal of Machine Learning Re-
search*, 6:1705–1749, 2005.

Beck, A. and Teboulle, M. A fast iterative shrinkage-
thresholding algorithm for linear inverse problems.
*SIAM J. Imag. Sci.*, 2(1):183–202, March 2009. ISSN
1936-4954. doi: 10.1137/080716542.

Beckmann, M. A continuous model of transportation.
*Econometrica*, 20:643–660, 1952.

Bigot, J. and Klein, T. Consistent estimation of a
population barycenter in the Wasserstein space.
*Preprint arXiv:1212.2562*, 2012.
Bonneel, Nicolas, Rabin, Julien, Peyré, Gabriel, and Pfister, Hanspeter. Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, pp. 1–24, 2013.

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

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

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

Cuturi, Marco and Doucet, Arnaud. Fast computation of Wasserstein barycenters. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, 2014.

Indyk, Piotr and Thaper, Nitin. Fast image retrieval via embeddings. In *3rd International Workshop on Statistical and Computational Theories of Vision (at ICCV)*, 2003.

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

Ling, H. and Okada, K. An efficient earth mover’s distance algorithm for robust histogram comparison. 29(5):840–853, 2007.

Lowe, D.G. Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pp. 1150 –1157 vol.2, 1999.

Nesterov, Y. Gradient methods for minimizing composite objective function. CORE Discussion Papers 2007076, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), Sep 2007. URL http://ideas.repec.org/p/cor/louvco/2007076.html.

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

Oliva, A. and Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42 (3):145–175, 2001.

Pele, Ofir and Werman, Michael. Fast and robust earth mover’s distances. In *ICCV'09*, 2009.

Pennington, Jeffrey, Socher, Richard, and Manning, Christopher D. Glove: Global vectors for word representation. *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014)*, 12, 2014.

Rabin, Julien, Peyré, Gabriel, Delon, Julie, and Bernot, Marc. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision*, volume 6667 of *Lecture Notes in Computer Science*, pp. 435–446. Springer, 2012.

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

Salton, Gerard, Wong, Anita, and Yang, Chung-Shu. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613 –620, 1975.

Sandler, Roman and Lindenbaum, Michael. Non-negative matrix factorization with earth mover’s distance metric. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1873–1880. IEEE, 2009.

Shirdhonkar, Sameer and Jacobs, David W. Approximate earth movers distance in linear time. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8. IEEE, 2008.

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

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

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

Zen, Gloria, Ricci, Elisa, and Sebe, Nicu. Simultaneous ground metric learning and matrix factorization with earth mover’s distance. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pp. 3690–3695. IEEE, 2014.

Zou, Will Y, Socher, Richard, Cer, Daniel M, and Manning, Christopher D. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*, pp. 1393–1398, 2013.