Convex transport potential selection with semi-dual criterion

Adrien Vacher, François-Xavier Vialard

Convex transport potential selection with semi-dual criterion

Adrien Vacher
LIGM, Univ. Gustave Eiffel, CNRS
INRIA
adrien.vacher@u-pem.fr

François-Xavier Vialard
LIGM, Univ. Gustave Eiffel, CNRS
francois-xavier.vialard@u-pem.fr

December 13, 2021

Abstract

Over the past few years, numerous computational models have been developed to solve Optimal Transport (OT) in a stochastic setting, where distributions are represented by samples. In such situations, the goal is to find a transport map that has good generalization properties on unseen data, ideally the closest map to the ground truth, unknown in practical settings. However, in the absence of ground truth, no quantitative criterion has been put forward to measure its generalization performance although it is crucial for model selection. We propose to leverage the Brenier formulation of OT to perform this task. Theoretically, we show that this formulation guarantees that, up to a distortion parameter that depends on the smoothness/strong convexity and a statistical deviation term, the selected map achieves the lowest quadratic error to the ground truth. This criterion, estimated via convex optimization, enables parameter and model selection among entropic regularization of OT, input convex neural networks and smooth and strongly convex nearest-Brenier (SSNB) models. Last, we make an experiment questioning the use of OT in Domain-Adaptation. Thanks to the criterion, we can identify the potential that is closest to the true OT map between the source and the target and we observe that this selected potential is not the one that performs best for the downstream transfer classification task.

1 Introduction

Optimal transport (OT) is a tool to compare probability distributions that has found numerous applications ranging from economics (Galichon, 2016; Chiappori et al., 2010), unsupervised learning (Sim et al., 2020), shape matching (Feydy et al., 2017), NLP (Chen et al., 2019; Alvarez-Melis and Jaakkola, 2018) and biology (Schiebinger et al., 2019; Tong et al., 2020). In its dual form, OT is a linear maximization problem on functions, which are called potentials, subject to a cost constraint. When the cost is chosen to be quadratic, the solutions of this problem are convex and their gradient provide optimal maps that transport one distribution onto the other. In a significant part of the OT applications, the transport map itself is the object of interest. For instance in Domain-Adaptation, the source distribution is transported on the target (Courty et al., 2017), for color transfer one color histogram is transported on the other (Rabin et al., 2014) and in biology, the RNA cell expression profile is interpolated in time using OT maps (Schiebinger et al., 2019). Over the past few years, many models and computational methods (Cuturi, 2013; Genevay et al., 2016; Seguy et al., 2018; Bonneel and Coeurjolly, 2019; Vacher et al., 2021) were proposed and implemented to estimate these optimal transport maps. Under regularity assumptions, some of these models were shown to accurately estimate the original transport map provided the models use optimal parameters (Pooladian and Niles-Weed, 2021; Manole et al., 2021). When such results exist,

either the parameters to use are explicit but they are impractical as they rely on generic worst-case bounds, either they involve unavailable constants. To the best of our knowledge, no quantitative criterion has yet been devised to calibrate the parameters of OT models and later discriminate between calibrated models. The setting we are interested in is standard in statistical and machine learning applications, in which probability measures are only accessible via samples in Euclidean spaces. The goal is to recover, for the quadratic cost, a potential chosen among different models/parameters that has good generalization properties and if possible, the closest to the unknown ground truth. For achieving this task, we put forward the use of the semi-dual functional of OT that we now define.

The semi-dual (Brenier) objective. The quantitative criterion that we propose is the so-called *semi-dual* Brenier objective of OT. It is a convex functional on the space of functions $f : \mathbb{R}^d \to \mathbb{R}$ and is defined for μ, ν two probability measures on the Euclidean space by, denoting $\langle \cdot, \cdot \rangle$ the pairing between Radon measures and continuous functions,

$$J_{\mu,\nu}(f) \coloneqq \langle f, \mu \rangle + \langle f^*, \nu \rangle \,, \tag{1}$$

where f^* is the Fenchel-Legendre transform of f given by

$$f^*(y) \coloneqq \sup_{x \in \mathbb{R}^d} x^\top y - f(x) \,. \tag{2}$$

Note that for a general cost c, this new objective can be obtained by replacing one potential by its c-transform in the Kantorovitch dual formulation. However, this case is particularly attractive in the case where f is convex since the pointwise computation of f^* is a concave maximization problem, which can be parallelized for different values of y. Indeed, if ν is a finite sum of m Dirac masses at points y_j, Formula (1) requires the computation of $f^*(y_j)$ which are m independent concave optimization problems.

Related works. The problem of evaluating OT models was recently studied by Korotin et al. (2021). They proposed to generate synthetic ground truth optimal maps using an input convex neural network. Then, they calibrate various OT models on these ground truth OT maps and they compare the performance of each calibrated OT model by measuring the natural L^2 distance between the estimated map and the ground truth. Their paper gives an interesting perspective on comparing current OT models. However, their setup requires the knowledge of the ground truth to calibrate the OT models. This limitation can be overcome for models providing convex potentials as shown in our work.

The use of the Fenchel-Legendre transform can be found in the pioneering paper Brenier (1991). It can be shown that this new formulation retains more convexity than the Kantorovitch formulation. On the theoretical side, this gain was then leveraged for uses as diverse as sharp bounds for the problem of statistical map estimation (Hütter and Rigollet, 2021) or quantitative stability results of the transport map with respect to the measures (Delalande and Merigot, 2021). On the numerical side, since the Fenchel-Legendre transform has linear cost on a grid (Fast Legendre Transform), the semi-dual is used to design efficient numerical algorithms in low dimension (Jacobs and Léger, 2020). In the machine learning community, the semi-dual was proposed by Taghvaei and Jalali (2019) to estimate convex transport maps parametrized by Input Convex Neural Networks (Amos et al., 2017). When n is the sample size, they noticed that instead of the classical $O(n^2)$ complexity of OT, this new formulation leads to an $O(n)$ complexity per iteration as it only requires n-independent computations of the Legendre transform.

Our contributions. Even if the Brenier semi-dual is frequently met in the OT literature, it is the first time that this objective is used for model and parameter selection. For this purpose, we also use the extra convexity of the semi-dual that allows us to prove quantitative bounds for the selection criterion.

The goal of our paper is to answer the following question: given $(f_1, \cdots, f_p)$, p convex potentials, how to select the one that minimizes the quadratic error,

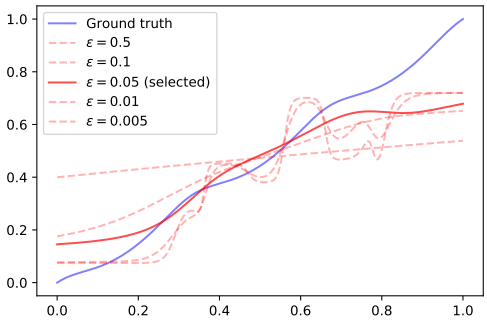

Figure 1: Entropic OT map selection with Brenier criterion in dimension 8 on 1024 samples for the Tensorized experiment (see Sec.5.2).

denoting ∇f_i the gradient of f_i,

$$e_\mu(f_i) = \int_x \|\nabla f_i(x) - T_0(x)\|^2 d\mu(x)\,, \qquad (3)$$

where T_0 is the true, unknown, OT map from μ to ν ?

1. We prove that the potential that minimizes the Brenier objective on a test set is the one that minimizes the error e_μ up to a multiplicative factor that depends on the smoothness of the potential and up to an additive statistical deviation term.

2. Entropic regularization of OT is a popular approach and we show this model can be efficiently adapted to our setup. We propose an efficient, GPU-friendly numerical scheme to compute the semi-dual Brenier objective.

3. We showcase on three synthetic experiments, among three different models, that in practice, the best transport potential is indeed selected and we nearly observe a monotone behavior of the error with respect to the value of semi-dual.

4. We perform a Domain-Adaptation experiment suggesting that, perhaps counter-intuitively, the best mapping from the source to the target does not generate the best performance on the classification task.

5. In addition to the previous results, we also provide the first publicly available implementation of the SSNB model proposed by Paty et al. (2020). In comparison with their algorithm, ours has better scaling properties thanks to an explicit SOCP reformulation of their original QCQP formulation.

Assumptions and notations In this paper X, Y are compact subsets of $\mathbb{R}^d$, μ and ν are probability measures over X and Y respectively with their n-samples empirical counterparts $\hat{\mu}$, $\hat{\nu}$. We shall denote by $\mathrm{supp}(\mu)$, $\mathrm{supp}(\nu)$ the support of μ and ν respectively.

2 Optimal transport and Brenier formulation

In its dual formulation, the OT problem optimizes over a pair of continuous functions, called *Kantorovitch* potentials (ϕ, ψ) subject to a cost constraint as

$$\mathrm{OT}_c(\mu,\nu) = \sup_{(\phi,\psi)} \langle \phi, \mu\rangle + \langle \psi, \nu\rangle + \iota(\phi \oplus \psi \leq c)\,, \quad \text{(D)}$$

where $\phi \oplus \psi$ is defined as $(\phi \oplus \psi)(x, y) = \phi(x) + \psi(y)$ and ι is the convex indicator function. When c is the Euclidean squared distance, we simply denote it by OT. In this case, if one of the two measures has density w.r.t. the Lebesgue measure, Brenier's theorem (Brenier, 1991) shows that a unique optimal map sending μ to ν exists and is given by the gradient of a convex function. If one further assumes regularity of the underlying densities and convexity of the support of the distributions, the optimal map is even smoother than simply being continuous, as detailed next.

Theorem 1 (Caffarelli (2000)). *Assume that μ and ν have C^1 densities bounded from below and above. If X, Y are compact and convex sets in $\mathbb{R}^d$, then, defining the Brenier potentials $(f, g) = (\frac{\|.\|^2}{2} - \phi, \frac{\|.\|^2}{2} - \psi)$,*

f and g are C^2 convex functions such that

$$\begin{cases} \nabla f_{\#}(\mu) = \nu \\ \nabla g_{\#}(\nu) = \mu \,, \end{cases} \tag{4}$$

where $T_{\#}(\eta)$ is the pushforward of the distribution η by the map T defined as $T_{\#}(\eta)(A) = \eta(T^{-1}(A))$ for all Borel A.

The semi-dual Brenier objective for convex potential selection. In the context of statistical OT, the transport Kantorovitch potentials $\hat{\phi}, \hat{\psi}$ are usually estimated using the dual formulation (D). To compare the obtained potentials, one may be tempted to simply evaluate the Kantorovitch linear objective on a test set

$$K_{\hat{\mu},\hat{\nu}}(\hat{\phi}, \hat{\psi}) = \langle \hat{\phi}, \hat{\mu} \rangle + \langle \hat{\psi}, \hat{\nu} \rangle \,, \tag{5}$$

where $\hat{\mu}, \hat{\nu}$ represent independent samplings of μ, ν. Note that the dense inequality constraint $\iota(\hat{\phi} \oplus \hat{\psi} \leq c)$ over $X \times Y$ has to be fulfilled. However, in numerous OT models, the learned potentials $(\hat{\phi}, \hat{\psi})$ usually do not respect this cost constraint. For instance, in the entropic regularization of OT the constraint is "loosely" satisfied on the train set since it replaces the hard inequality constraint by the soft penalization $\varepsilon \langle e^{\frac{c - \phi \oplus \psi}{\varepsilon}}, \hat{\mu} \otimes \hat{\nu} \rangle$. It is possible though to remove the cost constraint from the objective in order to evaluate candidate potentials. Rewriting the Kantorovich dual with the Brenier potentials gives

$$\begin{aligned} \mathrm{OT}(\mu, \nu) = & \inf_{(f,g)} \langle \frac{\|.\|^2}{2} - f, \mu \rangle + \langle \frac{\|.\|^2}{2} - g, \nu \rangle \\ & \text{s. t. } f(x) + g(y) \geq x^\top y \\ = & \langle \frac{\|.\|^2}{2}, \mu + \nu \rangle - \inf_{(f,g)} \langle f, \mu \rangle + \langle g, \nu \rangle \\ & \text{s. t. } f(x) + g(y) \geq x^\top y \,, \end{aligned} \tag{6}$$

where the optimization is done on $f, g \in C^0$, the space of continuous functions. The last inequality implies that $g(y) \geq \sup_x x^\top y - f(x)$ for every x which shows that we can replace g by f^*, the Fenchel-Legendre transform of f. Therefore, up to moment terms, we get the *semi-dual* Brenier formulation

$$\inf_f J_{\mu,\nu}(f) = \langle f, \mu \rangle + \langle f^*, \nu \rangle \,. \tag{7}$$

This new nonlinear objective gains in convexity with respect to the Kantorovitch formulation (see Section 3) and now, whenever f is strongly convex, $J_{\mu,\nu}(f)$ is finite and can be efficiently computed on discrete measures using standard convex optimization algorithms. Hence, if we restrict ourselves to convex f, possibly regularized with the addition of a small quadratic term, we are provided with a natural and well-behaved selection criterion: the potential that minimizes $J_{\hat{\mu},\hat{\nu}}$. We show in the next section that, thanks to the extra convexity, the minimization of this objective coincides, up to a stochastic term, with the minimization of the quadratic error $e_\mu(f) = \int_X \|\nabla f(x) - T_0(x)\|^2 d\mu$.

3 Potentials selection

In this section, we show our main result on the selection of potentials. We first start with a standard result which shows that the semi-dual formulation gains convexity. More precisely, we show that it is upper-bounded and lower-bounded by the quadratic error e_μ. A similar result can be found in Hütter and Rigollet (2021)[Proposition 10] but we give here a slightly sharper and more general version. When no confusion is possible, we shall from now on denote $J_{\mu,\nu}$ by J.

Proposition 1. *Assuming that an optimal convex potential f_0 such that $T_0 = \nabla f_0$ pushes μ onto ν exists, then if f is a γ-strongly convex C^1 function with M-Lipschitz gradient, we have*

$$\frac{1}{2M} e_\mu(f) \leq (J(f) - J(f_0)) \leq \frac{1}{2\gamma} e_\mu(f) \,, \tag{8}$$

where $e_\mu(f) = \|\nabla f - T_0\|^2_{L^2(\mu)}$.

We give a short proof hereafter for completeness.

Proof. The semi-dual functional can be rewritten as $J(f) = \langle f, \mu \rangle + \langle f^* \circ T_0, \mu \rangle$. Now, the Fenchel inequality on f gives for every couple $(x, y) \in X \times Y$ $y^\top x \leq f(x) + f^*(y)$, and equality holds for $y = \nabla f(x)$. To simplify notations, let us denote $T(x) = \nabla f(x)$. We get

$$f(x) + f^*(T(x)) = x^\top T(x) \,. \tag{9}$$

Denoting f_0 an optimal potential, the optimality condition of the OT problem also gives the equality $f_0(x) + f_0^*(T_0(x)) - x^\top T_0(x) = 0\,, \forall x \in Supp(\mu)$ the support of μ and by integration $J(f_0) = \int x^\top T_0(x) d\mu$. Therefore, we have

$$J(f) - J(f_0) = \int f(x) + f^*(T_0(x)) - x^\top T_0(x) d\mu\,.$$

By (9), using $f(x) = -f^*(T(x)) + x^\top T(x)$ in the previous equation and recalling that $\nabla f^*(T(x)) = x\ \forall x \in X$, we get

$$J(f) - J(f_0) = \int D_{f^*}(T_0(x), T(x)) d\mu\,, \tag{10}$$

where D_{f^*} is the Bregman divergence associated with f^* defined as $D_h(x,y) = h(x) - h(y) - (x-y)^\top \nabla h(y)$. Recall that if f is a C^1 convex function with a M-Lipschitz gradient, then f^* is $\frac{1}{M}$-strongly convex. Conversely, if f is a C^1 γ-strongly convex function, then f^* has $\frac{1}{\gamma}$-Lipschitz gradient. This implies the results:

$$\begin{cases} J(f) - J(f_0) \geq \frac{1}{2M} \int \|T_0(x) - T(x)\|^2 d\mu \\ J(f) - J(f_0) \leq \frac{1}{2\gamma} \int \|T_0(x) - T(x)\|^2 d\mu\,. \end{cases} \tag{11}$$

□

Thanks to these bounds, we prove our main result on convex map selection in a stochastic setting.

Proposition 2. *Let $(f_1, \cdots, f_p)$ be p potentials and $(\hat{\mu}, \hat{\nu})$ be the n-samples empirical counterparts of (μ, ν). Let i_0 be the index of the map that minimizes the empirical semi-dual, $i_0 = \arg\min_i \hat{J}(f_i)$ and similarly $i_1 = \arg\min_i e_\mu(f_i)$. If f_{i_0}, f_{i_1} are γ-strongly convex and M-smooth and if an OT map between μ and ν exists, then for all $0 < \delta < 1$ we have with probability at least $1 - \delta$*

$$e_\mu(f_{i_0}) \leq \frac{M}{\gamma} e_\mu(f_{i_1}) + 8MC\sqrt{\frac{\ln(4/\delta)}{2n}}\,, \tag{12}$$

where $C = \max(C_{i_0}, C_{i_1})$ with $C_i = \max(\|f_i\|_{X,o}, \|f_i^*\|_{Y,o})$ and $\|.\|_{Z,o}$ is defined as $\|g\|_{Z,o} = \sup_{z\in Z} g(y) - \inf_{z\in Z} g(y)$.

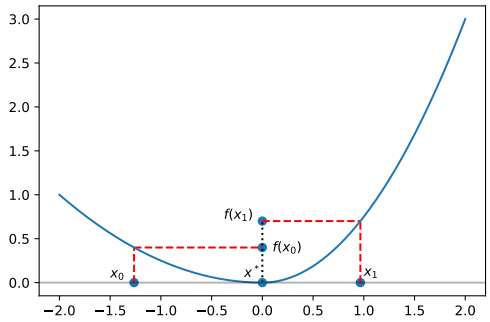

Figure 2: $f(x_0) < f(x_1)$ but x_0 is further from the minimum than x_1 and in particular we have the distortion $(x_0 - x^*)^2 \approx \frac{M}{\gamma}(x_1 - x^*)^2$.

Before starting the proof, we highlight the fact that the $\frac{M}{\gamma}$ distortion factor is sharp even in a non-stochastic setting. As shown by Figure 2, consider the function $f : x \mapsto \frac{\gamma}{2}x^2$ if $x < 0$ and $f(x) = \frac{M}{2}x^2$ if x is positive. f is indeed M-smooth, γ-strongly convex and attains its minimum in $x^* = 0$. For $\epsilon \to 0$, if we take the points $x_0 = -\frac{1}{\sqrt{\gamma}} + \epsilon$ and $x_1 = \frac{1}{\sqrt{M}} + \epsilon$, we have that $f(x_0) < f(x_1)$, and yet $\frac{(x_0 - x^*)^2}{(x_1 - x^*)^2} \to \frac{M}{\gamma} > 1$.

Proof. We begin with splitting $\hat{J}(f_{i_0}) - J(f_*)$ in non-stochastic and stochastic terms

$$\hat{J}(f_{i_0}) - J(f_*) = J(f_{i_0}) - J(f_*) + \hat{J}(f_{i_0}) - J(f_{i_0})\,. \tag{13}$$

Using Proposition 1, we get the lower bound

$$\hat{J}(f_{i_0}) - J(f_*) \geq \frac{1}{2M} e_\mu(f_{i_0}) + \hat{J}(f_{i_0}) - J(f_{i_0})\,. \tag{14}$$

By construction, f_{i_0} verifies for all $1 \leq i \leq p$

$$\begin{aligned} \hat{J}(f_{i_0}) - J(f_*) &\leq \hat{J}(f_i) - J(f_*) \\ &= J(f_i) - J(f_*) + \hat{J}(f_i) - J(f_i)\,. \end{aligned}$$

Picking $i = i_1$ and using Proposition 1, we obtain

$$\hat{J}(f_{i_0}) - J(f_*) \leq \frac{1}{2\gamma} e_\mu(f_{i_1}) + \hat{J}(f_{i_1}) - J(f_{i_1})\,. \tag{15}$$

Equations (14) and (15) give

$$e_\mu(f_{i_0}) \leq \frac{M}{\gamma} e_\mu(f_{i_1}) + 2M(\hat{J}(f_{i_1}) - J(f_{i_1})) \quad (16)$$
$$+ 2M(J(f_{i_0}) - \hat{J}(f_{i_0}))\,. \quad (17)$$

The Hoeffding lemma gives for all $t > 0$

$$\mathbb{P}(\langle f_i, \hat{\mu} - \mu\rangle \geq t) \leq \exp\left(-\frac{2nt^2}{\|f_i\|^2_{osc,X}}\right). \quad (18)$$

We place ourselves on the event

$$\begin{aligned} A =& (\langle f_{i_1}, \hat{\mu} - \mu\rangle \geq t) \cup (\langle f^*_{i_1}, \hat{\nu} - \nu\rangle \geq t) \\ & \cup (\langle f_{i_0}, \mu - \hat{\mu}\rangle \geq t) \cup (\langle f^*_{i_0}, \nu - \hat{\nu}\rangle \geq t)\,. \end{aligned} \quad (19)$$

We want to set $\mathbb{P}(A) \leq \delta$. By triangle inequality, we get the upper-bound

$$\mathbb{P}(A) \leq 4\exp\left(-\frac{2nt^2}{C^2}\right), \quad (20)$$

where $C = \max(C_{i_0}, C_{i_1})$ and C_i defined as $C_i = \max(\|f_i\|_{X,o}, \|f_i^*\|_{Y,o})$. Hence setting, $t = C\sqrt{\frac{\ln(4/\delta)}{2n}}$, we have with probability at least $1-\delta$

$$e_\mu(f_{i_0}) \leq \frac{M}{\gamma} e_\mu(f_{i_1}) + 8MC\sqrt{\frac{\ln(4/\delta)}{2n}}\,. \quad (21)$$

□

4 Sinkhorn potentials

The Sinkhorn model (Cuturi, 2013), defined as

$$S_\varepsilon(\mu,\nu) = \sup_{\phi,\psi\in C^0} \langle\phi,\mu\rangle + \langle\psi,\nu\rangle - \varepsilon\langle e^{\frac{\phi\oplus\psi - c}{\varepsilon}}, \mu\otimes\nu\rangle\,,$$

is very popular in the OT community. We show in this section that it indeed provides convex and smooth empirical potentials. Because of the lack of strong convexity, the semi-dual diverges on these empirical potentials. We show experimentally in Section 5 that using a small quadratic regularization does not degrade the performances of our selection method. Finally, we show how to efficiently compute the semi-dual on Sinkhorn potentials: from an algorithmic point of view, we show that they are *self-concordant* hence we can employ provably convergent second order schemes. From a numerical point of view, we provide GPU memory-friendly routines to compute their gradients and hessians.

A convex model. In order to compute the semi-dual, we first need to check whether the Brenier potentials associated with the Sinkhorn model are convex. Recall that the first-order optimality condition on the Sinkhorn potential ϕ_ε gives

$$\phi_\varepsilon(x) = -\varepsilon\log\left(\int_y e^{\frac{\psi_\varepsilon(y) - c(x,y)}{\varepsilon}} d\nu(y)\right). \quad (22)$$

Defining, for a quadratic cost $c(x,y) = \frac{\|x-y\|^2}{2}$, the associated Brenier potentials as $f_\varepsilon = \frac{\|.\|^2}{2} - \phi_\varepsilon(.)$, we obtain $f_\varepsilon(x) = \varepsilon\log(\int_y e^{\beta_\varepsilon(y)} e^{\frac{x^\top y}{\varepsilon}} d\nu(y))$, where we defined $\beta_\varepsilon(y) = e^{\frac{2\psi_\varepsilon(y) - \|y\|^2}{2\varepsilon}}$. This shows that the potentials $(f_\varepsilon, g_\varepsilon)$ are *Log-Sum-Exp* functions and in particular, they are convex. We draw the attention on the fact that we shall use (22) as it is classically done in the literature (Berman, 2018; Pooladian and Niles-Weed, 2021) to extend the empirical Sinkhorn potentials $(\hat{\phi}_\varepsilon, \hat{\psi}_\varepsilon) = \arg\min_{(\phi,\psi)} S_\varepsilon(\hat{\mu},\hat{\nu})$ that are originally solely defined on the samples $\hat{\mu}$ and $\hat{\nu}$. As we shall see later, this very particular parametric form can be proven useful in the numerical resolution of the Fenchel-Legendre transform. However, since Log-Sum-Exp are not strongly convex functions, we show below that these potentials need to be regularized.

Proposition 3. *Let* $(\hat{f}_\varepsilon, \hat{g}_\varepsilon) = (\frac{\|.\|^2}{2} - \hat{\phi}_\varepsilon, \frac{\|.\|^2}{2} - \hat{\psi}_\varepsilon)$. *If* ν *has continuous density with respect to the Lebesgue measure we have almost surely*

$$\langle \hat{f}^*_\varepsilon, \nu\rangle = +\infty\,. \quad (23)$$

The proof is left in Appendix. Obviously in such situation, the theoretical bound of Section 3 does not apply. Adding a small quadratic regularization of the form $\frac{\delta}{2}\|x\|^2$ to the potential f_ε makes the semi-dual value finite although it implies a bias on the selected potential. However, we show in Section 5 that it gives satisfying results in practice.

Self-concordant potentials. Now, even if the Legendre transform can be computed via a standard convex first order minimization algorithm, it will not be effective in practice. We show in Appendix that the Sinkhorn potentials are $O(\frac{1}{\varepsilon})$ smooth leading to a $O(\frac{1}{\delta\varepsilon})$ condition number on the problem, hence we need to employ second order methods. The flaw of these methods is that they require costly line-searches. However if the function has a *generalized self-concordant* structure, we can use a second order algorithm of the form

$$x_{k+1} = x_k - \alpha_k (\nabla^2 f(x_k))^{-1} \nabla f(x_k)\,, \tag{24}$$

where α_k is an explicit step size given in Sun and Tran-Dinh (2019)[Theorem 2] and that provably yields a super-linearly convergent algorithm.

Definition 1. *Let $\alpha > 0$ and f be a C^3 convex function. The function f is said to be (α, M_f) self-concordant if for all (x, u, v)*

$$|(\nabla^3 f(x)[v]u)^\top u| \leq M_f \|u\|_x^2 \|v\|_x^{\alpha-2} \|v\|^{3-\alpha}\,, \tag{25}$$

where $\|u\|_x^2 = (\nabla^2 f(x)u)^\top u$, $\nabla^3 f(x)$ is the tensor $(\frac{\partial^3 f}{\partial x_i x_j x_k})_{1\leq ijk\leq d}$ and for a tensor $T = (t_{ijk})_{1\leq i,j,k\leq n}$, $T[v] = \sum_{i=1}^p v_i T_i$ with T_i the matrix $(t_{ijk})_{1\leq j,k\leq n}$.

Informally, the self-concordance measures how fast the Hessian varies with respect to the metric it induces.

Proposition 4. *The Sinkhorn Brenier potential $\hat{f}_\varepsilon$ is $(2, \frac{D(\hat{\nu})}{\varepsilon})$ self-concordant where the diameter D is defined as $D(\hat{\nu}) = \sup_{y\in\hat{\nu}, z\in\hat{\nu}} \|y - z\|_2$.*

The proof is left in Appendix. Note that we can obtain the coarser non-stochastic bound $\frac{D(Supp(\nu))}{\varepsilon}$. For instance, if the distributions are contained in the unit cube, the self-concordant constant will be $\frac{\sqrt{d}}{\varepsilon}$. The Figure 3 shows that the second order scheme is much faster to compute the Fenchel conjugate on $n = 1000$ points, with learning rate $O(\frac{1}{\varepsilon})$ for the first order method.

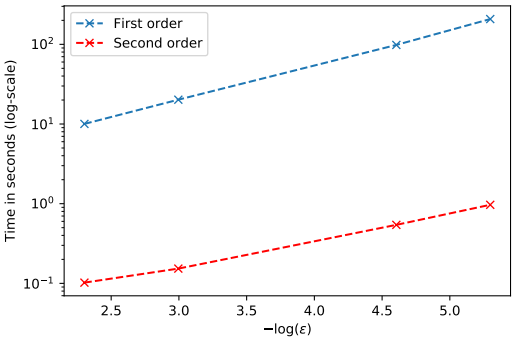

Figure 3: Fenchel Conjugate: 1st order vs 2nd order.

A GPU-friendly implementation Now that we can use the algorithm of Sun and Tran-Dinh (2019) with explicit constants to compute the Fenchel transform, we present in this paragraph a GPU-friendly implementation that parallelizes the updates across $\hat{\nu}$. Let $\hat{\nu} = (y_1, \cdots, y_n)$ and z_j the j-th current point of the j-th minimization problem

$$\min_z (\hat{f}_\varepsilon(z) + \frac{\delta}{2}\|z\|^2) - \langle z, y_j\rangle\,. \tag{26}$$

To design a GPU-friendly algorithm, we must first compute all the gradients $(\nabla f_\varepsilon(z_j))_{1\leq j\leq n} \in \mathbb{R}^{n\times d}$ simultaneously. The gradient is given by

$$\nabla f(z_j) = \frac{\sum_{i=1}^n y_i k_{ij}}{k_j}\,, \tag{27}$$

where $k_{ij} = \frac{1}{n} e^{\langle z_j, \frac{y_i}{\varepsilon}\rangle + \beta_\varepsilon(z_j)}$ and $k_j = \sum_{i=1}^n k_{ij}$. A naive computation would explicitly store the "kernel" k_{ij} leading to a $O(n^2)$ memory footprint. Instead we use the Keops library (Feydy et al., 2020) that symbolically computes k_{ij}. The same process is used for the $\sum_{i=1}^n y_i k_{ij}$ term. Conversely, we must compute the tensor of the Hessians $(\nabla^2 f_\varepsilon(z_j))_{1\leq j\leq n} \in \mathbb{R}^{n\times d\times d}$. As we show in Appendix, the Hessian is given by

$$\nabla^2 f_\varepsilon(z_j) = \frac{1}{\varepsilon}\left(\frac{\sum_{i=1}^n y_i y_i^\top k_{ij}}{k_j} - \nabla f_\varepsilon(z_j)\nabla f_\varepsilon^\top(z_j)\right).$$

Once again, we use the Keops library to reduce the memory footprint of $\sum_{i=1}^n y_i y_i^\top k_{ij}$ to $O(nd^2)$.

5 Numerical experiments

We first introduce two other transport models on which we perform our numerical experiments.

5.1 Other models

Input Convex Neural Network (ICNN). The convex Brenier potentials are modeled by $(f_{\theta_1}, g_{\theta_2})$ two ICNN. Then the model is trained using a minimax objective

$$\min_{\theta_1} \max_{\theta_2} \langle f_{\theta_1}, \hat{\mu}\rangle + \frac{1}{n}\sum_{i=1}^{n} \nabla g_{\theta_2}(y_i)^\top y_i - \langle f_{\theta_1}, (\nabla g_{\theta_2})_\#(\hat{\nu})\rangle \,.$$

The maximization aims at recovering $g_{\theta_2} \approx f^*_{\theta_1}$ and the minimization approximately solves the semi-dual. The implementation is based on the code of the authors[1]. Softplus activation layers were used instead of ReLu to obtain less degenerated maps. Note that since the weights θ_1, θ_2 are not controlled, we expect this model to provide lowly regular maps.

Smooth Strongly convex Nearest Brenier (SSNB). This model estimates the potential f by solving

$$\inf_{f \in \mathcal{F}_{(l,L)}} W_2((\nabla f)_\#(\hat{\mu}), \hat{\nu}) \,, \tag{28}$$

where $\mathcal{F}_{(l,L)}$ is the set of L-smooth, l-strongly convex functions. As opposed to the previous model, SSNB provides maps that are very regular (in a bi-Lipschitz sense). The algorithm is based on an alternate minimization scheme and the costly steps are Quadratically Constrained Quadratic Programs (QCQP) of the form

$$\begin{aligned} &\inf_{Z \in \mathbb{R}^{n\times(d+1)}} \frac{1}{2} Z^\top Q_0 Z + c_0^\top Z \\ &\text{s.t } \frac{1}{2} Z^\top Q_{ij} Z + c_{ij}^\top Z + r_{ij} \leq 0 \,, \end{aligned} \tag{29}$$

where $i \neq j$, $1 \leq i, j \leq n$ and Q_{ij}/Q_0 are sparse matrices. Indeed, one can straightforwardly implement this QCQP using CVXPY (Diamond and Boyd, 2016) but the resulting problem hardly scales up to $n = 100$ as a single iteration takes hours on a 120 GB RAM CPU. The main limitation is the overhead of the compilation, which is done at each iteration of the algorithm. Our implementation widely reduces the time per iteration to a few minutes for $n = 1024$. The details of the implementation are given in the Appendix; it relies on two key points: First, we explicitly implement the SOCP in MOSEK and in particular, the factorization of the Q_{ij} matrices is made "by hand". Second, the resulting cone constraints are compiled only once for a fixed value of (n, d) and the resulting problem is explicitly stored.

[1] https://github.com/AmirTag/OT-ICNN

5.2 The experiments

Synthetic XP. We compare the ability of the models to recover the ground truth transport map using the semi-dual criterion map for three different distributions in a medium dimension setting $d = 8$. In all three cases, the distribution μ is uniform on the cube $[0, 1]^d$ and ν is given by $(\nabla f)_\#(\mu)$ where f is a convex function ; in virtue of Brenier's theorem, $T = \nabla f$ is the ground truth OT map between μ and ν. The function f has 3 different forms

Quadratic: $f(x) = \frac{1}{2} x^\top Q x + x^\top b$ where $Q = O^\top D O + 0.25\,\mathrm{Id}$ where O is a randomly chosen orthogonal matrix, D is a random diagonal matrix whose entries are uniform in $[0, 1]$ and b is a random d-dimensional gaussian. This is a standard benchmark which simply aims at recovering a translation.

Tensorized: $T_0(x) = x + (6 - \cos(6\pi x) - 0.2)^{-1}$ and $T(x) = \sum_{k=1}^{d} T_0(x_k)$. The map to learn is more complex but has a low dimensional structure as it pushes independently each directions.

(Regularized) Log-Sum-Exp: $f(x) = t\,\mathrm{LSE}(\frac{C}{t}x + b) + \frac{\delta}{2}\|x\|^2$ where the matrix C is comprised of 10 centers uniformly chosen in $[-1, 1]^d$, the shift b is a random d-dimensional gaussian, the temperature t was fixed at 0.3 and the regularizer $\delta = 0.001$. Note that any convex function can be approximated by such a Log-Sum-Exp (Calafiore et al., 2020). However, because of this parametric structure, we expect the Sinkhorn model to be favored.

The training of the models, the semi-dual esti-

	ICNN		Sinkhorn		SSNB		ICNN/Sinkhorn/SSNB	
	$e_{\hat{\mu}}(f_{i_1})$	$e_{\hat{\mu}}(f_{i_0})$	$e_{\hat{\mu}}(f_{i_1})$	$e_{\hat{\mu}}(f_{i_0})$	$e_{\hat{\mu}}(f_{i_1})$	$e_{\hat{\mu}}(f_{i_0})$	$e_{\hat{\mu}}(f_{i_1})$	$e_{\hat{\mu}}(f_{i_0})$
Q	5.11	12.23 (16.13/48)	0.036	0.047 (1.93/5)	0.013	**0.014** (1.33/11)	0.013	0.014 (1.33/64)
T	2.74	2.74 (1.22/48)	0.059	0.119 (2.72/5)	0.006	**0.006** (1.0/11)	0.006	0.006 (1.0/64)
L	1.69	3.02 (17.11/48)	0.006	**0.006** (1.68/5)	0.16	0.17 (1.26/11)	0.006	0.006 (1.68/64)

Table 1: Potential Selection for best map recovery. Q, T and L stand for the Quadratic, Tensorized and Log-Sum-Exp experiments respectively. In bold the model that performed best after being calibrated with the semi-dual criterion. The last column corresponds to the performance of the model selected with the semi-dual criterion among the 3 calibrated ones. The numerator between brackets corresponds to the rank of the calibrated/selected potential with respect to the error $e_{\hat{\mu}}$: the closer to one, the better. The denominator corresponds to the number of potentials among which the chosen one was selected.

mation and the error Monte-Carlo approximation[2] $e_{\hat{\mu}}(f_i) = \int \|\nabla f_i(x) - T_0(x)\|^2 \mathrm{d}\hat{\mu}(x)$ are done with 3 independent batches of size $n = 1024$. Forty-eight combinations of hyperparameters were tested for the ICNN model, five for the Sinkhorn model and eleven for the SSNB model. More details on the hyperparameters and on the implementation are given in Appendix. For each experiment, the results were averaged on fifteen independent runs.

	Quad	Tensorised	Log-Sum-Exp
$e_{\mu}(f_{i_1})$	0.0104	0.0376	0.0005
$e_{\mu}(f_{i_0})$	0.0104	0.0376	0.0005
Rank	1.0/5	1.0/5	1.0/5

Table 2: Potential Selection with Sinkhorn Model with $n = 10000$ train/test/eval points. The numerator of the Rank corresponds to the rank of the potential calibrated with the semi-dual criterion with respect to the error $e_{\hat{\mu}}$: the closer to one, the better. The denominator corresponds to the number of potentials among which the chosen one was selected. We observe that the best candidate is always chosen by the semi-dual criterion.

The results are reported on Table 1. First the parameters of each model are calibrated with the semi-dual criterion (the three first columns of the Table). We denote i_0 the index of the minimizer of the semi-dual and i_1 the index of the potential that achieves the lowest $e_{\hat{\mu}}$ error. The numerator between brackets corresponds to the rank of the selected model with respect to the error $e_{\hat{\mu}}$; the closer to one the better. In particular, if $i_0 = i_1$ we obtain the rank 1. The denominator corresponds to the number of parameters: for instance, since we tested the Sinkhorn model with five different temperatures ε, we have a denominator of 5.

In the case of SSNB where the smoothness and strong convexity parameters are explicitly controlled, the best potential is almost always selected. In the case of the Sinkhorn model, the regularity decreases for small values of ε yet the selected potential remains in the top 40% for the Quadratic and Log-Sum-Exp experiments. Conversely, the regularity is not controlled in the ICNN model yet the selected potential remains in the top-tier for the Quadratic and Log-Sum-Exp experiments; as for the Tensorised experiment, the best potential is almost always selected.

Once the three models are calibrated with the semi-dual, we select one among them with the same criterion (last column of the Table); note the denominator between brackets is now equal to the sum of the number of parameters considered for each model. We observe that the criterion selects the best or nearly-best transport map. This shows that the Brenier criterion can be both used for calibration and selection. Figure 4 plots the semi-dual values against the error in the Log-Sum-Exp experiment and empirically suggests an even better behavior than best model selec-

[2] Concentrates in $O(\frac{1}{\sqrt{n}})$ toward the "true" error.

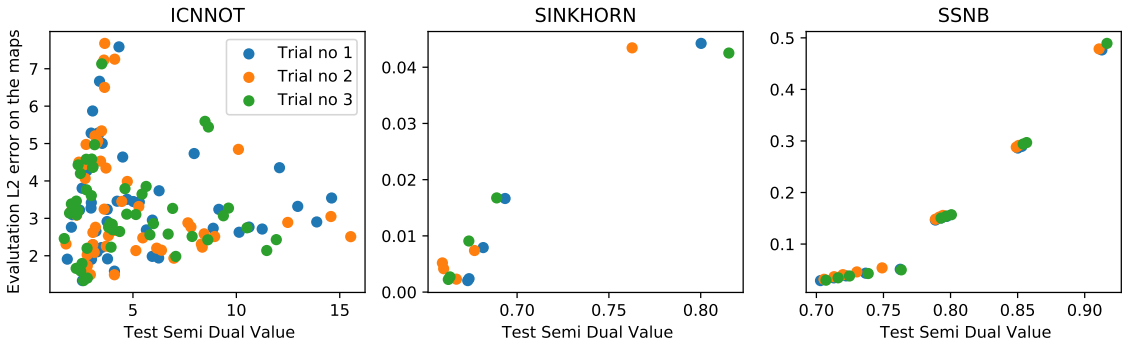

Figure 4: Empirical Semi-Dual against Quadratic Error on the Log-Sum-Exp experiment. Ideally, the error should strictly increase with the semi-dual; because of the distortion factor and the stochasticity stated in Proposition 2, we may not obtain this ideal behavior. We do obtain it for the SSNB model where the regularity is well controlled and nearly obtain it for the Sinkhorn model where we can at least control the gradient. The ICNNOT model behaves more poorly as expected since we do not control explicitly the regularity of the model.

tion. For the Sinkhorn and SSNB models, the error strictly increases with the semi-dual value, hence the semi-dual can not only select but can also directly rank the potentials with respect to the error e_μ. For the ICNN model, we do not observe the same monotone behavior but we still get a positive correlation between the error and the semi-dual value. This less consistent behavior can be explained once again by the lack of control on the regularity of the potentials given by the model.

Overall, when we compare the models after being calibrated with the semi-dual we observe that ICNN always has the poorest performance. We may not have chosen the hyperpameters and the network structures in the best possible way and the ground truth may not favour this model. The SSNB model performs better by almost an order of magnitude than Sinkhorn on the Quadratic and Tensorized experiments. Conversely, as expected, the Sinkhorn model is the best one on the Log-Sum-Exp experiment. In terms of computation time, the training of the SSNB model takes between one and three hours and between 30 minutes and one hour for the semi-dual computation on a 120 GB RAM CPU. ICNN and Sinkhorn take a few minutes and a few seconds respectively for the training and semi-dual computation on a RTX6000 GPU.

Thanks to the high scalabilty of Sinkhorn, we repeated those three experiments on larger batches with $n = 10000$ averaged on 10 runs. As shown on Table 2, Sinkhorn recovers the same behavior as SSNB with respect to the semi-dual. Not only the best parameter ε is always chosen but also, we show in the Appendix that in the Quadratic and Log-Sum-Exp setting, the error strictly increases with the semi-dual value.

Domain Adaptation. (DA) In DA, a map T is sought between a source distribution X_s with its known labels Y_s and a target distribution X_t with unknown labels. Then a classifier c is learned on $(T(X_s), Y_s)$ and is used to predict the unknown labels of the target as $c(X_t)$. In the work of Courty et al. (2017) and many others (Redko et al., 2019; Xu et al., 2020), the core assumption is that T should be close to the OT map between X_s, X_t. Question: is this assumption valid? Is the "true" OT map between X_s and X_t the one that will achieve the best knowledge transfer? Problem: among all proposed models, how to assess which map is the closest to the "true" *unknown* OT map? Using our criterion, we can now select the parameters and the model that

will be closest to the ground truth.

We use the Caltech-office dataset which is a set of images of objects from ten distinct categories coming from four different sources of various quality: objects found in the online Amazon catalog (A), objects whose pictures have been taken with a webcam (W), with a high resolution digital SLR camera (D) and the Caltech-256 dataset (C) which is comprised of Google images. We use all the nine distinct pairs as source/domain data. As in Courty et al. (2017), in order for the quadratic distance to be meaningful, we do not use the raw images but feed them to a Decaf (Donahue et al., 2014) network and extract the features of the last layer and we use a 1-NN as the classifier. The models and parameters we use are the same as in the Synthetic experiment. In our setting, the transport map T is learned on train sets X_s^{train}, X_t^{train} and the semi-dual is evaluated on test sets X_s^{test}, X_t^{test}, with 70% of the data for the train and 30% for the test.

The results are reported on Table 3. We denote by i_0 the index of the potential that minimizes the empirical semi-dual criterion and by i_1 the potential that achieves the highest accuracy. The numerator between bracket corresponds to the rank of the selected potential with respect to the accuracy obtained when the classifier is learned on $(\nabla f(X_s), Y_s)$; the closer to 1, the better and in particular, $\text{rank}(f_{i_1}) = 1$. We observe that the accuracy of the potential selected by the semi-dual is quite random. Worse, the potential that has the lowest accuracy is regularly selected by the semi-dual, even in the case of the SSNB model for which the Brenier criterion indicates very reliably the quality of the transport map. Hence we conclude that for DA, the best mapping for label transfer is not an optimal transport map, at least among our models. We remark that this conclusion is similar to the results of Korotin et al. (2021) and Stanczuk et al. (2021) who observed that the transport models which performed best for various ML tasks were not the ones that recover the sharpest OT maps.

6 Conclusion

The semi-dual Brenier formulation of quadratic OT provides us with a feasible criterion for convex po-

	ICNN		Sinkhorn		SSNB	
	$\text{acc}(f_{i_1})$	$\text{acc}(f_{i_0})$	$\text{acc}(f_{i_1})$	$\text{acc}(f_{i_0})$	$\text{acc}(f_{i_1})$	$\text{acc}(f_{i_0})$
A/C	0.41	0.34 (2/48)	0.84	**0.82** (3/5)	0.85	0.79 (10/11)
A/D	0.44	0.15 (33/48)	0.87	0.78 (4/5)	0.82	**0.8** (5/11)
A/W	0.36	0.07 (48/48)	0.78	**0.72** (3/5)	0.79	0.71 (9/11)
C/A	0.47	0.09 (44/48)	0.91	0.82 (5/5)	0.91	**0.88** (9/11)
C/D	0.65	0.27 (6/48)	0.9	0.8 (4/5)	0.88	**0.82** (10/11)
C/W	0.36	0.34 (3/48)	0.82	0.79 (2/5)	0.83	**0.83** (1/11)
D/A	0.5	0.47 (2/48)	0.91	0.78 (5/5)	0.91	**0.84** (11/11)
D/C	0.54	0.43 (2/48)	0.83	0.74 (5/5)	0.83	**0.75** (9/11)
D/W	0.52	0.28 (11/48)	0.96	0.85 (4/5)	0.99	**0.95** (8/11)
W/A	0.48	0.25 (17/48)	0.89	**0.78** (4/5)	0.87	0.77 (11/11)
W/C	0.4	0.2 (13/48)	0.77	0.73 (4/5)	0.78	**0.74** (10/11)
W/D	0.62	0.51 (2/48)	0.95	0.9 (3/5)	1.0	**1.0** (1/11)

Table 3: Potential Selection for Domain-Adaptation. The column $\text{acc}(f_{i_1})$ corresponds to the best (highest) accuracy and $\text{acc}(f_{i_0})$ corresponds to the accuracy of the potential selected with the Brenier criterion. On this Table, the potentials are ranked with respect to the accuracy; the closer to one, the better the classification. In bold, the highest accuracy after being calibrated with the semi-dual. For Domain-Adaptation the potential that is closest to an OT map does not yield the best accuracy.

tential selection. Provided the potentials are convex, this criterion can be computed numerically. Theoretically and experimentally, we showed that up to a sharp distortion parameter, the potential that minimizes the semi-dual is the one whose gradient minimizes the squared error to the ground truth map. Hence we believe that this criterion provides a fair and accurate procedure to benchmark convex OT models and solves the tricky question of hyperparameter tuning in the context of stochastic OT. Possible extensions could include the treatment of more general cost c and more general potentials such as non-smooth convex or non-convex potentials.

References

David Alvarez-Melis and Tommi Jaakkola. Gromov-wasserstein alignment of word embedding spaces. In *Conference on Empirical Methods in Natural Language Processing*, 2018.

Brandon Amos, Lei Xu, and J Zico Kolter. Input convex neural networks. In *ICML*, 2017.

Robert J Berman. The Sinkhorn algorithm, parabolic optimal transport and geometric monge-amp\ere equations. *Num. Math.*, 2018.

Nicolas Bonneel and David Coeurjolly. Spot: Sliced partial optimal transport. *SIGGRAPH*, 2019.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 1991.

Luis A Caffarelli. Monotonicity properties of optimal transportation. *Communications in Mathematical Physics*, 2000.

Giuseppe Carlo Calafiore, Stéphane Gaubert, and Corrado Possieri. Log-sum-exp neural networks and posynomial models for convex and log-log-convex data. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

Liqun Chen, Yizhe Zhang, Ruiyi Zhang, Chenyang Tao, Zhe Gan, Haichao Zhang, Bai Li, Dinghan Shen, Changyou Chen, and Lawrence Carin. Improving sequence-to-sequence learning via optimal transport. In *ICLR*, 2019.

Pierre-André Chiappori, Robert J. McCann, and Lars P. Nesheim. Hedonic price equilibria, stable matching, and optimal transport: Equivalence, topology, and uniqueness. *Economic Theory*, 2010.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013.

Alex Delalande and Quentin Merigot. Quantitative stability of optimal transport maps under variations of the target measure. *arXiv*, 2021.

Steven Diamond and Stephen Boyd. CVXPY: A python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 2016.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014.

Jean Feydy, Benjamin Charlier, François-Xavier Vialard, and Gabriel Peyré. Optimal transport for diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017.

Jean Feydy, Joan Glaunès, Benjamin Charlier, and Michael Bronstein. Fast geometric learning with symbolic matrices. *NeurIPS*, 2020.

Alfred Galichon. *Optimal transport methods in economics*. Princeton University Press, 2016.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic Optimization for Large-scale Optimal Transport. In *NeurIPS*, 2016.

Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 2021.

M. Jacobs and F. Léger. A fast approach to optimal transport: the back-and-forth method. *Numerische Mathematik*, 2020.

Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, Alexander Filippov, and Evgeny Burnaev. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. *arXiv*, 2021.

Tudor Manole, Sivaraman Balakrishnan, Jonathan Niles-Weed, and Larry Wasserman. Plugin estimation of smooth optimal transport maps, 2021.

François-Pierre Paty, Alexandre d'Aspremont, and Marco Cuturi. Regularity as regularization: Smooth and strongly convex brenier potentials in optimal transport. In *AISTATS*, 2020.

Giovanni Pistone and Henry P Wynn. Finitely generated cumulants. *Statistica Sinica*, 1999.

Aram-Alexandre Pooladian and Jonathan Niles-Weed. Entropic estimation of optimal transport maps, 2021.

Julien Rabin, Sira Ferradans, and Nicolas Papadakis. Adaptive color transfer with relaxed optimal transport. In *2014 IEEE International Conference on Image Processing (ICIP)*, 2014.

Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. Optimal transport for multi-source domain adaptation under target shift. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 2019.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 2019.

Vivien. Seguy, Bharath B. Damodaran, Remi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. In *ICLR*, 2018.

Byeongsu Sim, Gyutaek Oh, Jeongsol Kim, Chanyong Jung, and Jong Chul Ye. Optimal transport driven cyclegan for unsupervised learning in inverse problems. *SIAM Journal on Imaging Sciences*, 2020.

Jan Stanczuk, Christian Etmann, Lisa Maria Kreusser, and Carola-Bibiane Schönlieb. Wasserstein gans work because they fail (to approximate the wasserstein distance). *arXiv*, 2021.

Tianxiao Sun and Quoc Tran-Dinh. Generalized self-concordant functions: a recipe for newton-type methods. *Mathematical Programming*, 2019.

Amirhossein Taghvaei and Amin Jalali. 2-wasserstein approximation via restricted convex potentials with application to improved training for gans. *arXiv*, 2019.

Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *ICML*, 2020.

Adrien Vacher, Boris Muzellec, Alessandro Rudi, Francis Bach, and Francois-Xavier Vialard. A dimension-free computational upper-bound for smooth optimal transport estimation. In *COLT*, 2021.

Renjun Xu, Pelen Liu, Liyan Wang, Chao Chen, and Jindong Wang. Reliable weighted optimal transport for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Additional proofs

Proof of Proposition 3

Proof. Recall that the Fenchel-Legendre of a standard Log-Sum-Exp function $\mathrm{LSE}(x) = \log(\sum_{i=1}^n e^{x_i})$ is given by

$$\mathrm{LSE}^*(y) = \sum_{i=1}^n y_i \log(y_i) + \iota(y \in \mathcal{S}_n) \tag{30}$$

$$= -\operatorname{Ent}(y) + \iota(y \in \mathcal{S}_n)\,, \tag{31}$$

where $\mathcal{S}_n$ is the probability simplex. More generally, defining $\mathrm{LSE}_b(x) = \log(\sum_{i=1}^n e^{x_i+b_i})$, using the fact that $f^*(\cdot+\tau) = f^*(\cdot) - \tau^\top \cdot$, we have

$$(\mathrm{LSE}_b)^*(y) = -\operatorname{Ent}(y) - b^\top y + \iota(y \in \mathcal{S}_n)\,. \tag{32}$$

At the optimum, for empirical measures $\hat{\mu} = \frac{1}{m}\sum_{i=1}^m \delta_{x_i}, \hat{\nu} = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ the empirical Sinkhorn Kantorovitch potentials $(\hat{\phi}_\varepsilon, \hat{\psi}_\varepsilon)$ are linked as

$$\hat{\phi}_\varepsilon(x) = -\varepsilon \log\left(\frac{1}{n}\sum_{i=1}^n e^{2\frac{2\hat{\psi}_\varepsilon(y_i) - \|x-y_i\|^2}{2\varepsilon}}\right), \tag{33}$$

hence the Sinkhorn Brenier potential f_ε can be written as

$$f_\varepsilon(x) = \varepsilon\, \mathrm{LSE}_{b_\varepsilon}(C_\varepsilon x)\,, \tag{34}$$

where $C_\varepsilon = (\frac{y_i}{\varepsilon})_{1\le i\le n} \in \mathbb{R}^{n\times d}$ and $b_{\varepsilon,n} = (\frac{2\psi_\varepsilon(y_i) - \|y_i\|^2}{2\varepsilon} - \log(n))_{1\le i\le n} \in \mathbb{R}^n$. Now recall that

- $(\varepsilon f(\cdot))^* = \varepsilon f^*(\frac{\cdot}{\varepsilon})$.
- $\forall z,\ (f(A.))^*(z) = \inf_{Ay=z} f^*(y)$.

Hence we can deduce

$$f_\varepsilon^*(y) = \varepsilon \inf_{\substack{C_1\Delta = y \\ \Delta\in\mathcal{S}_n}} -\operatorname{Ent}(\Delta) - \Delta^\top b_{\varepsilon,n} + \iota(\Delta \in \mathcal{S}_n)\,.$$

In particular if f_ε^* is evaluated outside the convex hull of $\hat{\nu}$, it is infinite. Since ν has continuous density, there almost surely exists (y_0, r), $r > 0$ such that $B(y_0, r) \subset Supp(\nu)$ and $B(y_0, r) \cap Conv(\hat{\nu}) = \emptyset$. In particular, almost surely

$$\langle f_\varepsilon^*, \nu\rangle = +\infty\,. \tag{35}$$

□

Proof of Proposition 4

The proof is largely inspired from an article on the online blog of Francis Bach[3].

Since the 2-self-concordance is scaling invariant, we shall simply prove that $f(x) = \mathrm{LSE}_b(C.)$ is $(2, D(C))$ self-concordant with $b \in \mathbb{R}_+^n$, $C \in \mathbb{R}^{n\times d}$ the matrix whose rows are centers $(c_i)_{1\le i\le n}$ and $D(C) = \max_{ij} \|c_i - c_j\|$.

[3] https://francisbach.com/self-concordant-analysis-for-logistic-regression/

Proof. Defining the (non-normalized) distribution $\mu = \frac{1}{n}\sum_{i=1}^n b_i \delta_{c_i}$, we can remark that f is the normalizing factor of the conditional exponential distribution

$$h(c|x) \propto e^{c^\top x} d\mu(c) \tag{36}$$
$$= e^{c^\top x - f(x)} d\mu(c)\,. \tag{37}$$

The gradient of f is given by

$$\nabla f(x) = \frac{\int c e^{c^\top x} d\mu(c)}{\int e^{c^\top x} d\mu(c)} \tag{38}$$
$$= \mathbb{E}_h(c)\,, \tag{39}$$

and using the results of Pistone and Wynn (1999), we have for higher order derivatives

$$\nabla^p f(x) = \mathbb{E}_h(\otimes_{j=1}^p (c - \nabla f(x)))\,, \tag{40}$$

where for a vector $v \in \mathbb{R}^d$, $\otimes_{j=1}^p v$ is a tensor V_p in $\mathbb{R}^{d^p}$ whose entries are $(v_{i_1,\cdots,i_p})$. In particular, applying the formula for $p = 3$ and denoting $H = (c - \nabla f(x)) \otimes (c - \nabla f(x))$

$$\nabla^3 f(x) = \mathbb{E}_h[(c - \nabla f(x)) \otimes H]\,. \tag{41}$$

Using the linearity of the expectation, we have

$$|(\nabla^3 f(x)[v]u)^\top u| = |\mathbb{E}_h[(c - \nabla f(x))^\top v \times (Hu)^\top u]| \tag{42}$$
$$\leq \mathbb{E}_h[|(c - \nabla f(x))^\top v| \times |(Hu)^\top u|]\,. \tag{43}$$

Since $\nabla f(x) \in Conv(C)$, we have in particular that $\|c - \nabla f(x)\| \leq D(C)$. Furthermore since H is a positive matrix, we obtain the following upper-bound

$$|(\nabla^3 f(x)[v]u)^\top u| \leq D(C)\|v\|\mathbb{E}_h[(Hu)^\top u] \tag{44}$$
$$\leq D(C)\|v\|(\nabla^2 f(x)u)^\top u\,. \tag{45}$$

□

Sinkhorn Brenier potentials are $\frac{1}{\varepsilon}$-smooth

Proof. Using the notations from above, the Sinkhorn Brenier empirical potentials are of the form $f_\varepsilon = \varepsilon\,\mathrm{LSE}_{b_{\varepsilon,n}}(C_\varepsilon .)$. Using the formulas from the previous proof, we simply have to bound $H_{c,x} = (c - \nabla f(x)) \otimes (c - \nabla f(x))$

$$u^\top H_{c,x} u = (u^\top (c - \nabla f(x)))^2 \tag{46}$$
$$\leq \|u\|_2^2 \|c - \nabla f(x)\|_2^2\,. \tag{47}$$

Since $\nabla f(x) \in Conv(\frac{\hat{\nu}}{\varepsilon})$ and $c \in Supp(\frac{\hat{\nu}}{\varepsilon})$, we deduce that $\|H_{c,x}\|_{op} \leq \frac{D^2(\hat{\nu})}{\varepsilon^2}$, where $\|.\|_{op}$ is the spectral norm. In particular $\|\nabla^2 f(x)\|_{op} \leq \frac{D^2(\hat{\nu})}{\varepsilon}$. Hence, if we regularize the Sinkhorn Brenier potential with $\delta\frac{\|.\|^2}{2}$, we obtain a $O(\frac{1}{\delta\varepsilon})$ condition number. □

Miscellaneous

SSNB algorithm

For $l < L$, the SSNB model is defined as

$$\inf_{f \in \mathcal{F}_{l,L}} W_2^2((\nabla f)_\#(\mu), \nu)\,, \tag{48}$$

where $\mathcal{F}_{l,L}$ is the set of l-strongly convex, L-smooth functions. For empirical potentials $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\hat{\nu} = \frac{1}{m}\sum_{i=1}^m \delta_{y_i}$, the authors propose to solve the non-convex problem (48) in an alternate fashion: for a fixed $f \in \mathcal{F}_{l,L}$, they estimate the transport coupling $(P_{ij}) \in \mathbb{R}^{n\times m}$ from $(\nabla f)_\#(\hat{\mu})$ to $\hat{\nu}$ by solving the associated linear program (or an entropic approximation) and then, once the coupling is fixed, they estimate f (pointwise on $\hat{\mu}$) by solving

$$\begin{aligned}&\min_{(z_1,\cdots,z_n)\in\mathbb{R}^{n\times d}, u\in\mathbb{R}^n} \sum_{ij} P_{ij}\|z_i - y_j\|_2^2\\ &\text{subject to } u_i \geq u_j + z_j^\top(x_i - x_j) + \frac{1}{2(1-l/L)}\left(\frac{1}{L}\|z_i - z_j\|^2 + \frac{1}{l}\|x_i - x_j\|_2^2 - \frac{2l}{L}(z_j - z_i)^\top(x_i - x_j)\right),\end{aligned} \tag{49}$$

where $z_i = \nabla f(x_i)$ and $u_i = f(x_i)$. The problem above is a convex Quadratically Constrained Quadratic Problem and can be numerically solved with CVXPY for instance. However, when such an option is chosen the $n(n-1)$ constraints must be computed at each iterations which induces a large overhead. Instead, we reformulate this problem as a standard linear conic problem of the form $Ax - b \in \mathcal{K}$, with $\mathcal{K}$ a fixed cone to be compiled only once.

From QCQP to SOCP First we show how to reformulate a (convex) QCQP without equality constraints into an SOCP. The standard formulation of a QCQP is

$$\begin{aligned}\inf_x \quad & \frac{1}{2}x^\top Q_0 x + c_0^\top x\\ \text{s. t. } & \frac{1}{2}x^\top Q_i x + c_i^\top x + r_i \leq 0, \quad i = 1,\cdots,p\,.\end{aligned} \tag{50}$$

Introducing the slack variables $(t_0, t_1, \cdots, t_p) = \frac{1}{2}(x^\top Q_0 x, x^\top Q_1 x, \cdots, x^\top Q_p x)$, we re-write the problem as

$$\begin{aligned}\inf_{x,t} \quad & t_0 + c_0^\top x\\ \text{s. t. } & t_i + c_i^\top x + r_i = 0, \quad i = 1,\cdots,p\\ & t_i \geq \frac{1}{2}x^\top Q_i x, \quad i = 0,\cdots,p\,.\end{aligned} \tag{51}$$

Decomposing Q_i as $Q_i = F_i^\top F_i$ with F_i having p rows, the constraint $t_i = \frac{1}{2}x^\top Q_i x$ becomes $(1, t_i, F_i x) \in \mathcal{Q}_r^{d+2}$, where $\mathcal{Q}_r^{d+2}$ is the rotated $(d+2)$-dimensional Lorentz cone defined as

$$\mathcal{Q}_r^{d+2} = \{(x_1, x_2, \cdots, x_{d+2}) \text{ s.t. } 2x_1x_2 \geq \sum_{k=1}^{d} x_{i+2}^2\}\,. \tag{52}$$

We obtain a MOSEK-friendly formulation of the QCQP as

$$\begin{aligned} &\inf_{x,t} \quad t_0 + c_0^\top x \\ &\text{s. t. } t_i + c_i^\top x + r_i = 0, \quad i = 1, \cdots, p \\ &\qquad (1, t_i, F_i x) \in \mathcal{Q}_r^{d+2}, \ i = 0, \cdots, p\,, \end{aligned} \tag{53}$$

which has the form $Ax - b \in \mathcal{K}$ where $\mathcal{K}$ is a fixed product of Lorentz cone whose number and dimensions solely depend on n and d in the case of SSNB. Hence we can compile $\mathcal{K}$ only once for fixed (n, d), which allows us to considerably reduce the overhead.

Decomposition of Q_{ij} In the SSNB model the symmetric positive matrices $Q_{ij} \in \mathcal{S}^+_{n(d+1)}(\mathbb{R})$ are defined up to a common scaling parameter as

$$\begin{cases} q_{kl} = 1 \text{ if } k = l \in \{di, \cdots (d+1)i\} \cup \{dj, \cdots (d+1)j\} \\ q_{kl} = -1 \text{ if } l = k + dj, k \in \{di, \cdots (d+1)i\} \\ q_{kl} = -1 \text{ if } k = l + dj, l \in \{di, \cdots (d+1)i\}\,. \end{cases} \tag{54}$$

The matrix Q_{ij} is factorized as $F_{ij}^\top F_{ij}$ with $F_{ij} \in \mathbb{R}^{d \times n(d+1)}$ defined as

$$\begin{cases} f_{kl} = 1 \text{ if } l = k + di, k \in \{1, \cdots, d\} \\ f_{kl} = -1 \text{ if } l = k + dj, k \in \{1, \cdots, d\}\,. \end{cases} \tag{55}$$

Models hyperparameters

ICNN We used a 3-layers ICNN with softplus activations. The number of hidden neurons was chosen in $\{64, 128, 256\}$, the soft convexity penalty for the potential g and the matching moment/variance penalty were both chosen in $\{0, 0.001, 0.01, 0.1\}$. As recommended by the authors, the batch size was set to 60, the number of epochs was set to 60, the number of inner iterations to approximate the conjugate was set to 25 and the learning rate is initially set to 1e-4 and is then divided by 2 every 2-epochs.

To compute the semi-dual, we regularized the potential f by adding $\frac{\delta}{2}\|x\|^2$ with $\delta =$ 1e-3. The numerical optimization was done with SciPy with a stopping condition set to 0.001 ; for a lower stopping criterion, the minimization would not converge.

Sinkhorn The temperature ε was chosen in $\{0.5, 0.1, 0.05, 0.01, 0.005\}$. We stopped the training when the optimality conditions are almost met

$$\begin{cases} \langle\ |\phi_\varepsilon(.) + \varepsilon \log(\int_y e^{\frac{\psi_\varepsilon(y) - c(.,y)}{\varepsilon}} d\hat{\nu}(y))|, \hat{\mu}\rangle \leq \text{1e-5} \\ \langle\ |\psi_\varepsilon(.) + \varepsilon \log(\int_x e^{\frac{\phi_\varepsilon(y) - c(x,.)}{\varepsilon}} d\hat{\mu}(x))|, \hat{\nu}\rangle \leq \text{1e-5}\,. \end{cases} \tag{56}$$

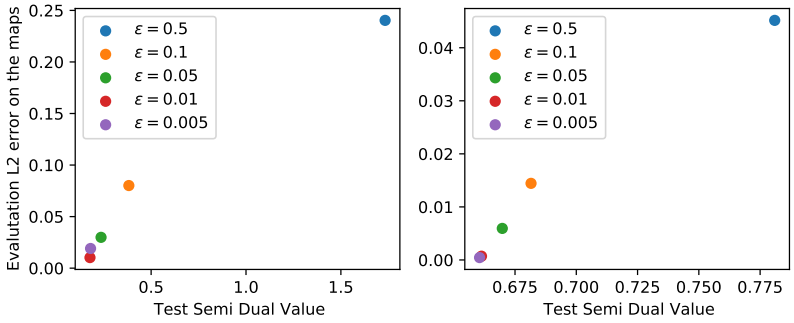

Figure 5: Empirical Semi-Dual against Quadratic Error on the Quadratic and Log-Sum-Exp experiments for the Sinkhorn model, $n = 10000$ and $d = 8$.

The resulting Sinkhorn Brenier potential $\hat{f}_\varepsilon$ is regularized with $\frac{\delta}{2}\|x\|^2$, $\delta = 0.001$. When the semi-dual is computed on a point y_i, the stopping criterion is given by

$$\|\nabla \hat{f}_\varepsilon(z_t) - y_i\| \leq \text{1e-5}\,, \tag{57}$$

where z_t is the current point of the optimization at time step t.

SSNB The strong convexity parameter l is chosen in $\{0.2, 0.5, 0.7, 0.9\}$ and the smoothness parameter L is chosen in $\{0.2, 0.5, 0.7, 0.9, 1.2\}$ with $l < L$. The number of iterations in the alternate minimization is set to 10. The conjugate is computed with a first order scheme with learning rate $\frac{1}{2L}$ and is stopped with the same criterion as above.

Additional Experiment Sinkhorn

Increasing n We run 10 times the Quadratic and Log-Sum-Exp experiments with the Sinkhorn model but with the train/test/eval sets of size $n = 10000$. The results are reported on Figure 6. Just as for SSNB, the semi-dual can accurately rank the potentials according to their error $e_\mu(f_i) = \int \|\nabla f_i(x) - T_0(x)\|_2^2 d\mu(x)$ where T_0 is the ground truth OT map.