AN ELEMENTARY INTRODUCTION TO ENTROPIC REGULARIZATION AND PROXIMAL METHODS FOR NUMERICAL OPTIMAL TRANSPORT

FRANÇOIS-XAVIER VIALARD

ABSTRACT. These notes contains the material that I presented to the CEA-EDF-INRIA summer school about numerical optimal transport. All the methods presented hereafter rely on convex optimization, so we start with a fairly basic introduction to convex analysis and optimization. Then, we present the entropic regularization of the Kantorovich formulation and present the now well known Sinkhorn algorithm, whose convergence is proven in continuous setting. Then, we present the linear convergence rate of this algorithm with respect to the Hilbert metric. The second numerical method we present use the dynamical formulation of optimal transport proposed by Benamou and Brenier which is solvable via non-smooth convex optimization methods.

1. INTRODUCTION

These notes are based on [Cuturi and Peyré, 2019]. For the convergence of the Sinkhorn algorithm, the proof is inspired by the proof in [Berman, 2017]. Most of the results on entropic regularization can be found in [Cuturi and Peyré, 2019]. The last results on Sinkhorn divergence are based on [Feydy et al., 2018]. For the numerical methods on the dynamical formulation, we rely on [Benamou and Brenier, 2000, Cuturi and Peyré, 2019, Papadakis et al., 2014, Chizat et al., 2018].

2. A GLIMPSE AT CONVEX ANALYSIS AND OPTIMIZATION

In the following, we choose to consider the setting of Hilbert spaces instead of the more general non-reflexive Banach spaces to benefit from the additional scalar product structure. However, optimal transport needs the more general case to include the case of Radon measures.

2.1. Usual definitions.

Definition 1. Let $C \subset E$ be a subset of the Banach space E, C is convex if for all $x, y \in C$, the segment $[x, y]$ is contained in C.

Of course the definition makes sense on a vector space but we need a topology on E for the Hahn-Banach theorem.

Definition 2. A function $f : E \mapsto [-\infty, \infty]$ is convex if its epigraph defined as

$$\operatorname{epi}(f) \overset{\text{def.}}{=} \{(x, y) : y \geq f(x)\} \subset E \times \mathbb{R} \tag{2.1}$$

is convex. The domain of f is $\operatorname{dom}(f) \overset{\text{def.}}{=} \{x : f(x) < +\infty\}$.

The function f is said proper if there exists $x_0 \in E$ such that $f(x_0) < +\infty$ and if f never takes the value $-\infty$. If f is proper, the definition of convexity reduces to the usual definition $f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y)$ for every couple $x, y \in E$ and $t \in [0, 1]$. Last, f is said strictly convex if the previous inequality is strict for $t \in]0, 1[$.

We want the function to be defined on the completed real line $[-\infty, \infty]$ in order to include constraints in the optimization problem.

Definition 3. A function $f : E \to \mathbb{R}$ is said lower semi-continuous (lsc) if for every $x_n \to x$

$$f(x) \leq \lim_{n \to \infty} f(x_n) \,. \tag{2.2}$$

Example 1. *Let $C \subset E$ be a set. We denote by $\iota_C : E \mapsto \mathbb{R}$ the indicator function of C defined as*

$$\iota_C(x) = \begin{cases} 0 \text{ if } x \in C\,, \\ +\infty \text{ otherwise.} \end{cases} \tag{2.3}$$

It convex iff C is convex, proper iff C is non-empty and lsc iff C is closed. This example is important in order to formulate constraint optimization problems as unconstrained optimization. More precisely, we mean

$$\min_{x \in C} f(x) = \min_{x \in C} f(x) + \iota_C(x)\,. \tag{2.4}$$

A direct consequence of the definition, we have the following fact,

Proposition 2 (Sup of convex function is convex)**.** *Let $f_i : E \to \mathbb{R}$ be convex functions indexed by a set I. Then, $\sup_{i \in I} f_i$ is a convex function.*

As a result of the Hahn-Banach theorem,

Proposition 3 (Closed + convex → weakly closed)**.** *A closed (for the strong topology) convex set is also closed for the weak topology (which differs in infinite dimension).*

An important property that is constantly used and is a consequence of Hahn-Banach theorem is

Proposition 4. *A convex lsc proper function is equal to the supremum of its affine minorants.*

To get a more quantitative description of this affine minorant, we need the definition of convex conjugate. Hereafter, we consider the case where E, E^* is a dual pair. For instance, when E is a Hilbert space or a finite dimensional space $E = E^*$. Optimal transport needs the more general case; Indeed, if X is a compact domain in $\mathbb{R}^d$, $E = C(X, \mathbb{R})$ is a Banach space when endowed with the sup norm and $E^* = \mathcal{M}(X)$ is the set of Radon measures.

Definition 4 (Convex conjugate)**.** Let $f : E \mapsto \mathbb{R}$ be a function. The convex conjugate $f^* : E^* \mapsto \mathbb{R}$ is defined as

$$f^*(p) = \sup_{x \in E} \langle p, x \rangle - f(x)\,. \tag{2.5}$$

Proposition 5. *Let $f : E \mapsto \mathbb{R}$ be a function, then f^{**} is the greatest lsc convex function below f. And, if f is convex lsc proper, then $f^{**} = f$.*

We now give the definition of the subgradient of a convex function which is the generalization of the gradient.

Definition 5 (Subgradient)**.** Let $f : E \to \mathbb{R}$ be a convex function and $x \in E$. The subgradient of f at point x is the set of elements in E^* defined by

$$\partial f(x) \overset{\text{def.}}{=} \{p \in E^* : f(y) \geq f(x) + \langle p, y - x \rangle \text{ for all } y \in E\}\,. \tag{2.6}$$

Remark 1. *If f is continuous at point x_0 then the subgradient at this point is non-empty, and also at every point in the interior of* $\mathrm{dom}(f)$*. The subdifferential can be empty at some points. In general, if E is a complete Banach space and f is convex lsc and proper, the set of points where ∂f is non-empty is dense in* $\mathrm{dom}(f)$*.*

Proposition 6. *The definition of subgradient implies, exchanging the order of x, y in the inequality* (2.6) *and adding the two inequalities*

$$\langle \partial f(x) - \partial f(y), x - y \rangle \geq 0\,, \tag{2.7}$$

with a little abuse of notations since $\partial f(x)$ and $\partial f(y)$ denote any element in these sets.

Proposition 7 (Legendre-Fenchel identity)**.** *Let f be a convex function. Then, the three statements are equivalent*

- $f(x) + f^*(p) = \langle p, x \rangle\,,$
- $p \in \partial f^*(x),$
- $x \in \partial f(p).$

Remark 2. *If f and f^* are differentiable, then the Legendre-Fenchel identity simply says that $\nabla f \circ \nabla f^* = \mathrm{Id}_{E^*}$ and $\nabla f^* \circ \nabla f = \mathrm{Id}_E$, which is sometimes a useful property to manipulate optimality formulas.*

Definition 6 (Strong convexity)**.** Let $\lambda > 0$ be a positive real. A convex function f is λ strongly convex if the function $x \mapsto f(x) - \frac{\lambda}{2}\|x\|^2$ is convex.

Proposition 8 (Strong convexity of f and smoothness of f^*)**.** *A convex function f is λ strongly convex iff f^* is C^1 with Lipschitz gradient with constant $1/\lambda$. Also, the subgradient satisfies*

$$\langle \nabla f^*(x) - \nabla f^*(y), x - y\rangle \geq \lambda \|\nabla f^*(x) - \nabla f^*(y)\|^2 \,, \tag{2.8}$$

∇f is a co-coercive monotone operator.

Definition 7 (Gradient flow and (explicit) gradient descent)**.** Let $f : H \mapsto \mathbb{R}$ be a C^1 function. The gradient flow associated with f is

$$\dot{x} = -\nabla f(x)\,, \tag{2.9}$$

with initial value $x(0) = x_0 \in H$.

A time-discrete counterpart is constant step size gradient descent, for $\tau > 0$,

$$x_{k+1} = x_k - \tau \nabla f(x_k) \tag{2.10}$$

Proposition 9. *If f is convex and C^1 with Lipschitz gradient of constant L, then the explicit gradient descent converges if $\tau < 2/L$ under the additional assumptions that f bounded below with bounded level sets.*

Proof. Only assuming f C^1 with Lipschitz gradient of constant L, implies that

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + L/2\|y - x\|^2\,, \tag{2.11}$$

and that the sequence of values $f(x_k)$ is decreasing since for $y = x_{k+1}$ and $y = x_k$, one has

$$f(x_{k+1}) \leq f(x_k) + \tau\langle \nabla f(x_k), \nabla f(x_k)\rangle + L\tau^2/2\|\nabla f(x_k)\|^2 \tag{2.12}$$

$$\leq \tau(-1 + L\tau/2)\|\nabla f(x_k)\|^2\,. \tag{2.13}$$

Therefore, if $\tau < 2/L$, $f(x_{k+1}) < f(x_k)$. If $(x_k)_{k\in\mathbb{N}}$ has an accumulation, which can be obtained under mild assumptions on the function f (as mentioned for instance bounded level sets in $\mathbb{R}^d$), then this accumulation point is a critical point of f. If f is convex, it is a global minimum and the sequence converges to this accumulation point since the map $x \mapsto x - \tau\nabla f(x)$ can be proven to be a weak contraction and thus the distance to this accumulation point is decreasing. □

If the objective function f is not C^1 with gradient L Lipschitz, it is possible to try to apply implicit gradient descent instead of explicit which iterates $x_{k+1} = x_k - \tau\nabla f(x_k)$.

Definition 8 (Implicit gradient descent and variational formulation)**.** The implicit gradient scheme with constant step size gradient descent, for $\tau > 0$,

$$x_{k+1} = x_k - \tau\nabla f(x_{k+1})\,. \tag{2.14}$$

This time-discrete scheme has a variational formulation,

$$x_{k+1} = \arg\min \frac{1}{2\tau}\|x - x_k\|^2 + f(x)\,, \tag{2.15}$$

which is uniquely defined if the function f is convex, proper and lsc (in this case, f has an affine minorant and the minimized function is coercive).

Proposition 10. *The so-called Moreau-Yosida regularization of f is $f_\tau(y) \overset{\text{def.}}{=} \min_x \frac{1}{2\tau}\|x - y\|^2 + f(x)$ and it is C^1 with $1/\tau$ Lipschitz gradient. The explicit gradient scheme for f_τ is the implicit gradient scheme for f and consequently, the implicit gradient descent converges independently of the choice of τ.*

Definition 9. Let f be a convex function, proper and lsc. The proximal operator is defined as

$$\operatorname{prox}_{\tau f}(x) = \arg\min_y \frac{1}{2\tau}\|x-y\|^2 + f(x)\,. \tag{2.16}$$

As said above, $\operatorname{prox}_{\tau f}(x)$ is uniquely defined and satisfies

$$\operatorname{prox}_{\tau f}(x) - x + \tau\partial f(x) \ni 0\,. \tag{2.17}$$

The notation $(\mathrm{Id} + \tau\partial f)^{-1}x = \operatorname{prox}_{\tau f}(x)$ will be used.

In particular, if it is reasonably cheap to compute the proximal operator of f, then the implicit gradient descent $x_{k+1} = \operatorname{prox}_{\tau f}(x_k)$ can be used. Such functions are called simple. Therefore, it is interesting to know that computing the proximal map of a function is as difficult as computing the proximal map of its convex conjugate.

Proposition 11. *Let f be a convex, proper and lsc function. Then, it holds*

$$x = \operatorname{prox}_{\tau f}(x) + \tau\operatorname{prox}_{\frac{1}{\tau}f^*}(\frac{1}{\tau}x)\,, \tag{2.18}$$

known as Moreau's identity.

Let us be interested in the following optimization problem,

$$\min_x f(x) + g(x)\,, \tag{2.19}$$

where f is simple function and g is a C^1 function with L Lipschitz gradient. At a critical point x_*, one has

$$f(x) + g(x) \leq f(x) + g(x_*) + \langle \nabla g(x_*), x - x_*\rangle + \frac{L}{2}\|x - x_*\|^2\,, \tag{2.20}$$

and therefore, it is natural to minimize the right-hand side which gives the composition of a proximal operator and a gradient step for g, since $\langle \nabla g(x_*), x - x_*\rangle + \frac{L}{2}\|x - x_*\|^2 = \frac{1}{2}\|\frac{1}{L}\nabla g(x_*) + x - x_*\|^2$,

$$\operatorname{prox}_{1/Lf} \tag{2.21}$$

We are now interested in the minimization problem

$$\min_x f(Kx) + g(x)\,, \tag{2.22}$$

where K is a bounded linear operator, f and g are convex, lsc and proper functions. In order to present the primal-dual algorithms, we now compute the dual problem associated to (2.22).

$$\min_x \max_p \langle p, Kx\rangle - f^*(p) + g(x) \geq \max_p \min_x \langle p, Kx\rangle - f^*(p) + g(x) \tag{2.23}$$

$$\geq \max_p -g^*(-K^*p) + f^*(p)\,, \tag{2.24}$$

Equality between the l.h.s and r.h.s. is satisfied under mild assumptions. In the case of non-reflexive Banach space, we recall a useful theorem in convex analysis, the Fenchel-Rockafellar theorem. We recall the notion of topologically paired spaces, E, E^* if

Theorem 12 (Fenchel-Rockafellar). *Let (E, E^*) and (F, F^*) be two topological dual pairs, $L : E \mapsto F$ be a continuous linear map and denote $L^* : F^* \mapsto E^*$ its adjoint. Let $f : E \mapsto \mathbb{R}$ and $g : F \mapsto \mathbb{R}$ be two proper, convex and lower semicontinuous functions. Under the following condition if there exists $x \in \mathrm{Dom}(f)$ such that g is continuous at Ax, the following equality holds*

$$\sup_{x\in E} -f(-x) - g(Lx) = \min_{p\in F^*} f^*(L^*p) + g^*(p)\,. \tag{2.25}$$

In case there exists a maximizer $x \in E$, then there exists $p \in F^*$ such that $Lx \in \partial g^*(p)$ and $L^*p \in \partial f(-x)$.

Note that the conclusion of the theorem has a dissymmetry, the minimum on the right-hand side being attained. Let us give an example of application with standard optimal transport: We consider a compact domain $X \subset \mathbb{R}^d$, $\rho_1, \rho_2 \in \mathcal{M}_1(X)$ two probability measures. On the space $X \times X$, we consider the space of nonnegative Radon measures.

3. Entropic regularization of optimal transport

The Kantorovich formulation of optimal transport aims at minimizing a linear function over the simplex $\mathcal{S}_{n,m}$ of probability vectors on $\mathbb{R}^{n \times m}$ defined by

$$\mathcal{S}_{n,m} = \{\pi_{ij} \in \mathbb{R}_+^{n\times m} \,:\, \sum_{i=1}^{n}\sum_{j=1}^{m} \pi_{ij} = 1\}\,. \tag{3.1}$$

Namely, denoting $\langle \cdot, \cdot \rangle$ the L^2 scalar product on $\mathbb{R}^{n\times m}$,

$$\mathrm{OT}(\rho_1, \rho_2) = \min \langle \pi(i,j), c(i,j) \rangle \text{ such that } \sum_j \pi_{i,j} = \rho_1(i) \text{ and } \sum_i \pi_{i,j} = \rho_2(j) \forall i, j\,. \tag{3.2}$$

This linear programming problem has complexity $O(N^3)$ which is clearly infeasible for large N, N being $\max(n,m)$. Moreover, as a linear programming problem the resulting cost $\mathrm{OT}(\rho_1, \rho_2)$ is not differentiable (everywhere) with respect to ρ_1, ρ_2.

Entropic regularization provides us with an approximation of optimal transport, with lower computational complexity and easy implementation.

Entropic regularization, in its continuous formulation, can actually be traced back to the seminal work of Schrödinger in the 20's, and has been rediscovered several times in different contexts. We refer to the book [Cuturi and Peyré, 2019] in which many historical references are cited. This section is motivated by the introduction of entropic regularization for the above mentioned reasons by Cuturi in [Cuturi, 2013]. In this paper, entropy penalty is added, as done in linear programming

$$\min_{\pi \in \Pi(\rho_1, \rho_2)} \langle \pi(i,j), c(i,j) \rangle - \varepsilon \,\mathrm{Ent}(\pi)\,, \tag{3.3}$$

where we denoted the set of admissible couplings by

$$\Pi(\rho_1, \rho_2) \overset{\text{def.}}{=} \{\pi \in \mathcal{S}_{n,m} \,:\, \sum_j \pi_{i,j} = \rho_1(i) \text{ and } \sum_i \pi_{i,j} = \rho_2(j) \forall i, j\}\,. \tag{3.4}$$

and the Shannon entropy, which is a strictly concave function

$$\mathrm{Ent}(\pi) \overset{\text{def.}}{=} -\sum_{i,j} \pi_{i,j}(\log(\pi_{i,j}) - 1)\,. \tag{3.5}$$

Therefore, problem (3.3) is strictly convex and by compactness of the simplex, there exists a unique solution. Due to the fact that $x\log(x)$ has infinite positive slope at 0, this minimizer satisfies that $\pi_{i,j} > 0$, and one can apply the first order optimality condition with constraints (KKT conditions), forming the Lagrangian associated with the problem

$$L(\pi, \lambda_1, \lambda_2) = \langle \pi(i,j), c(i,j) \rangle - \varepsilon \,\mathrm{Ent}(\pi) - \langle \lambda_1(i), \sum_j \pi_{i,j} - \rho_1(i) \rangle - \langle \lambda_2(j), \sum_i \pi_{i,j} - \rho_2(i) \rangle\,, \tag{3.6}$$

and we obtain taking variations

$$c(i,j) + \varepsilon \log(\pi_{i,j}) - \lambda_1(i) - \lambda_2(j) = 0\,. \tag{3.7}$$

This implies that the unique optimal coupling for entropic regularization is of the form

$$\pi_{ij} = e^{\lambda_1(i) + \lambda_2(j) - c(i,j)} = D_1 e^{-c(i,j)} D_2\,, \tag{3.8}$$

where D_1, D_2 denote the diagonal matrices formed by $e^{\lambda_1(i)}$ and $e^{\lambda_2(j)}$. In order to solve for λ_1, λ_2 or equivalently, D_1, D_2, the marginal constraints give information on D_1, D_2. The problem now takes a similar form to the matrix scaling problem,

Matrix Scaling Problem: *Let $A \in \mathbb{R}^{mn}$ be a matrix with positive coefficients. Find D_1, D_2 two positive diagonal matrices respectively in $\mathbb{R}^{n^2}$ and $\mathbb{R}^{m^2}$, such that $D_1 A D_2$ is doubly stochastic, that is sum along each row and each column is equal to* 1.

First, solutions are non-unique since, if (D_1, D_2) is a solution, then so is $(\lambda D_1, \frac{1}{\lambda} D_2)$ for every positive real λ. This problem can be solved in a cheap way by a simple iterative algorithm, known as Sinkhorn-Knopp algorithm, which simply alternates updating D_1 and D_2 in order to match the marginal constraints. This algorithm takes the form, denoting by $\mathbf{1}_n$ the vector of size n filled with the value 1. At iteration k, the algorithm consists in updating alternatively D_1 and D_2 via the formula,

$$\textbf{Sinkhorn algorithm:} \begin{cases} D_1^k = \mathbf{1}_n./(AD_2^{k-1}) \\ D_2^k = \mathbf{1}_m./(A^T D_1^k)\,, \end{cases} \tag{3.9}$$

where we denoted ./ the coordinatewise division. The convergence of this algorithm has been proven by Sinkhorn and Knopp. In our case, the corresponding iterations would take the form

$$\begin{cases} D_1^k = \rho_1./(e^{-c/\varepsilon} D_2^{k-1}) \\ D_2^k = \rho_2./([e^{-c/\varepsilon}]^T D_1^k)\,. \end{cases} \tag{3.10}$$

However, to recast entropic optimal transport as a particular instance of bistochastic matrix scaling, one simply replaces $e^{-c/\varepsilon}$ with $\mathrm{diag}(\rho_1) e^{-c/\varepsilon} \mathrm{diag}(\rho_2)$. Interestingly, it is easy to modify the variational formulation in order to obtain this matrix in the optimality equation and this motivates the following definition,

Definition 10 (Discrete Entropic OT)**.**

$$\mathrm{OT}_\varepsilon(\rho_1, \rho_2) \overset{\text{def.}}{=} \min_{\pi \in \Pi(\rho_1, \rho_2)} \langle \pi(i,j), c(i,j) \rangle + \varepsilon \,\mathrm{KL}(\pi \,|\, \rho_1 \otimes \rho_2)\,, \tag{3.11}$$

where $\mathrm{KL}(\rho \,|\, \mu)$ is the Kullback-Leibler divergence, or relative entropy between ρ and μ and it is defined in the discrete case as

$$\mathrm{KL}(\rho \,|\, \mu) \overset{\text{def.}}{=} \sum_i \rho(i) \left(\log(\rho(i)/\mu(i)) - 1\right)\,. \tag{3.12}$$

The main point of defining entropic regularization using mutual information is to define the problem on the whole space of measures, in particular containing discrete and continuous measures.

Remark 3. *A few remarks are in order:*

- *The Kullback-Leibler entropy is jointly convex as we will see below.*
- *Note that the regularization term is known as mutual information between two random variables X, Y of respective law ρ_1, ρ_2 and joint distribution π.*
- *Mutual information is not convex in all of its arguments but for instance in (π, ρ_1) or (π, ρ_2).*
- *The argmin of problems* (3.11) *and* (3.3) *are the same. The formulation* (3.3) *can be rewritten as using the $\mathrm{KL}(\pi \,|\, \mathbf{1} \otimes \mathbf{1})$ and a simple calculation show that the argmin is independent of the choice of the measures α, β in $\mathrm{KL}(\pi \,|\, \alpha \otimes \beta)$. Of course, the value of the minimization problem is changing.*
- *If the cost c is nonnegative, OT_ε is nonnegative since mutual information is nonnegative.*

As expected, the behaviour w.r.t large and small values of ε can be characterised.

Proposition 13 (Limit cases in ε)**.** *When ε goes to 0, the unique minimizer π_ε for $\mathrm{OT}_\varepsilon(\alpha,\beta)$ converges to the maximal entropy plan among the possible optimal transport plans for $\mathrm{OT}(\alpha,\beta)$.*

When ε goes to $+\infty$, the unique minimizer π_ε converges to $\alpha \otimes \beta$, i.e. the joint law encoding independence of marginals.

Proof. We refer to the proof in [Cuturi and Peyré, 2019]. □

As is usual for an optimization problem, the nonuniqueness case is rare althgouh it obviously happens in optimal transport: an example with sum of two Dirac masses can be easily built, for instance the vertices of a square. A sufficient condition for uniqueness of the transport plan is the case of Brenier's theorem where one of the two marginals is assumed absolutely continuous w.r.t. the Lebesgue measure. Nevertheless, the limit of the entropic plans converges to a unique solution which can be considered intuitively as the most "diffuse" solution.

3.1. **Convergence of Sinkhorn algorithm in the continuous setting.** As recalled in Fenchel-Rockafellar theorem 12, the supremum of the dual problem might not be attained. However, in standard optimal transport, existence of optimal potential can be proven by standard compactness arguments. In this paragraph, we show that similar arguments go through.

Coordinate ascent algorithm on a function of two variables $f(x,y)$ can be informally written as

$$y_{n+1} = \arg\max_y f(x_n, y) \tag{3.13}$$

$$x_{n+1} = \arg\max_x f(x, y_{n+1}) \,. \tag{3.14}$$

Sinkhorn algorithm is a coordinate ascent on the dual problem, which can be formulated as

Proposition 14 (Dual Problem)**.** *The dual problem reads $\sup_{u,v} D(u,v)$ where $u,v \in C^0(X)$ and*

$$D(u,v) = \langle u(x), \alpha(x)\rangle + \langle v(y), \beta(y)\rangle - \varepsilon\langle \alpha\otimes\beta, e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}} - 1\rangle \,. \tag{3.15}$$

It is strictly convex w.r.t. each argument u and v and strictly convex w.r.t. $u(x)+v(y)$. It is also Fréchet differentiable for the $(C^0, \|\cdot\|_\infty)$ topology. Last, $D(u,v) = D(u+C, v-C)$ for every constant $C \in \mathbb{R}$. If a maximizer exists, it is unique up to this invariance.

Proof. The strict convexity and smoothness follows from the strict convexity and smoothness of the exponential (the functional D is the sum of linear terms and an exponential term which is smooth w.r.t. its arguments in the $(C^0, \|\cdot\|_\infty)$ topology). By strict convexity, $u_{k+1} = \arg\min_u D(u, v_k)$ and $v_{k+1} = \arg\min_v D(u_{k+1}, v)$ are uniquely defined. The invariance is immediate to check and the strict convexity in $u(x)+v(y)$ gives that if two maximizers exist, (u_1,v_1) and (u_2,v_2) then, $u_1(x)+v_1(y) = u_2(x)+v_2(y)$ which implies $u_1(x)-u_2(x) = v_2(y)-v_1(y)$ and the existence of C such that $(u_1,v_1) = (u_2+C, v_2-C)$ follows. □

Proposition 15 (Sinkhorn algorithm on dual potentials)**.** *The maximization of $D(u,v)$ w.r.t. each variable can be made explicit, and the Sinkhorn algorithm is defined as*

$$u_{k+1}(x) = -\varepsilon \log\left(\int_X e^{\frac{v_k(y)-c(x,y)}{\varepsilon}}\,\mathrm{d}\beta(y)\right) (=: S_\beta(v_k)) \tag{3.16}$$

$$v_{k+1}(y) = -\varepsilon \log\left(\int_X e^{\frac{u_{k+1}(x)-c(x,y)}{\varepsilon}}\,\mathrm{d}\alpha(x)\right) (=: S_\alpha(u_{k+1}))\,. \tag{3.17}$$

Moreover, the following properties hold

- *$D(u_k, v_k) \leq\ D(u_{k+1}, v_k) \leq D(u_{k+1}, v_{k+1})$,*
- *The continuity modulus of u_{k+1}, v_{k+1} is bounded by that of $c(x,y)$.*
- *If $v_k - c$ (resp. $u_{k+1} - c$) is bounded by M on the support of β, then so is u_{k+1} (resp. v_{k+1}).*

Proof. We prove existence of maximizer by proving that there exists a critical point to the functional coordinatewise. The first part of the proposition follows from writing the first-order necessary condition, written as follows

$$1 - e^{u(x)} \int_X e^{\frac{v(y)-c(x,y)}{\varepsilon}} \,\mathrm{d}\beta(y) = 0 \text{ for } x\; \alpha \text{ a.e.} \tag{3.18}$$

which gives the definition of $S_\beta(v)$ (and by symmetry, the same result on S_α holds). Therefore, $S_\beta(v)$ is the unique maximizer of $u \mapsto D(u,v)$.

By definition of ascent on each coordinate, the sequence of inequalities is obtained directly.

For the second point, remark that the derivative of $\log(\sum_i \exp(x_i))$ w.r.t. x_j is $\frac{\exp(x_j)}{\sum_i \exp(x_i)}$ bounded by 1. Therefore, $x \mapsto \log \int_X e^{\frac{c(x,y)-v(y)}{\varepsilon}} \,\mathrm{d}\beta(y)$ is L-Lipschitz where L is the Lipschitz constant of c, and the modulus of continuity of u_{k+1}, v_{k+1} is thus bounded by that of c. The last point is a simple bound on the iterates. □

Remark 4 (Link with standard optimal transport)**.** *The Sinkhorn algorithm computes iterates u_{k+1}, v_{k+1} which are as smooth as its cost and the continuity modulus of the iterates is bounded. Thus, the situation is close to the usual c-transform of optimal transport: starting from potentials u, v, one can replace v by u^* while the dual value is non-decreasing. The c-transform being L-Lipschitz with a constant independent of u, the maximization can thus be performed on the space of L-Lipschitz functions (which take the value 0 at a given anchored point) which is compact by the Arzelà-Ascoli theorem. Therefore, proving the existence of optimal potentials.*

Proposition 16. *The sequence (u_k, v_k) defined by the Sinkhorn algorithm converges in $(C^0(X), \|\cdot\|_\infty)$ to the unique (up to a constant) couple of potentials (u,v) which maximize D.*

Proof. First, shifting the potentials by an additive constant, one can replace the optimization set by the couples (u,v) which have a uniformly bounded modulus of continuity and such that $u(x_0) = 0$ for a given $x_0 \in X$. The maximum of D is achieved at some couple (u_*, v_*) and this couple is unique up to an additive constant as written in Proposition 14.

Then, since (u_{k+1}, v_{k+1}) are uniformly bounded and have uniformly bounded modulus of continuity, one can extract, by the Arzelà-Ascoli theorem, a converging subsequence in the corresponding topology to $(\tilde{u}, \tilde{v})$. By continuity of D and monotonicity of the sequence of values, $D(\tilde{u}, S_\alpha(\tilde{u})) \leq D(S_\beta \circ S_\alpha(\tilde{u}), S_\alpha(\tilde{u})) = D(\tilde{u}, S_\alpha(\tilde{u}))$, where S is the Sinkhorn iteration. Therefore, the maximizer coordinatewise being unique, one has,

$$S_\beta(\tilde{v}) = \tilde{u} \tag{3.19}$$

$$S_\alpha(\tilde{u}) = \tilde{v}\,. \tag{3.20}$$

Formulas (3.19) (together with (3.19)) show that $(\tilde{u}, \tilde{v})$ is a critical point of D, thus being the maximizer.

□

In fact, a particularly important property used in the convergence proof is that the log-sum-exp function, also called log cumulant is 1-Lipschitz.

Proposition 17. *The LSE function* $\log \int \exp$ *is convex (but not strictly) and 1-Lipschitz. Also, one has, for α a probability measure whose support is not a singleton,*

$$\|S_\alpha(u_1) - S_\alpha(u_2)\|^\circ_\infty \leq \kappa \|u_1 - u_2\|^\circ_\infty \tag{3.21}$$

where $\kappa < 1$ and

$$\|f\|_{\circ,\infty} \overset{\text{def.}}{=} \|f - \int_X f(x)\,\mathrm{d}\alpha\|_{\infty,\alpha}\,. \tag{3.22}$$

where the sup norm is taken w.r.t. α.

Proof. The first part of the proposition is obvious and used in the proof of Proposition 15. More precisely, the 1-Lipschitz property can be actually obtained by using

$$|S_\alpha(u_1)(x) - S_\alpha(u_2)(x)| = \left| \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} S_\alpha(u_2 + t(u_1 - u_2))\,\mathrm{d}t \right| \tag{3.23}$$

$$\le \int_0^1 \left| \int_X (u_1 - u_2) \frac{e^{\frac{t(u_1-u_2)}{\varepsilon}}}{\int_X e^{\frac{t(u_1-u_2)}{\varepsilon}} e^{\frac{u_2-c(x,\cdot)}{\varepsilon}}\,\mathrm{d}\alpha} e^{\frac{u_2-c(x,\cdot)}{\varepsilon}}\,\mathrm{d}\alpha \right| \mathrm{d}t \tag{3.24}$$

$$\le \|u_1 - u_2\|_\infty\,. \tag{3.25}$$

The case of equality can happen if and only if $u_1 - u_2$ is α a.e. a constant. In such a case, $u_1 = u_2 + a$, $S_\alpha(u_1) = S_\alpha(u_2) + a$. Therefore, it is natural to consider $C^0(X)/\mathbb{R}$, the space of continuous functions up to an additive constant, which we endow with the norm defined in the proposition:

$$\|f\|_{\circ,\infty} = \|f - \int_X f(x)\,\mathrm{d}\alpha\|_{\infty,\alpha}\,. \tag{3.26}$$

Note that such an approach only applies to measures α whose support is not restricted to a single point (an obvious case for balanced optimal transport). Using the same arguments as above, one has, for $u_1 \neq u_2$

$$\|S_\alpha(u_1) - S_\alpha(u_2)\|_{\circ,\infty} \le \|S_\alpha(u_1) - S_\alpha(u_2)\|_\infty < \|u_1 - u_2\|_{\circ,\infty} \tag{3.27}$$

since the case of equality implies that $u_1 = u_2$. Refining the inequality above (3.23), one has

$$|S_\alpha(u_1)(x) - S_\alpha(u_2)(x)| \le \kappa \|u_1 - u_2\|_{\circ,\infty}\,, \tag{3.28}$$

where, κ is defined by optimization on the set

$$\mathcal{S} \overset{\text{def.}}{=} \{f \text{ of continuity modulus less than twice that of } c,\ \|f\|_{\circ,\infty} = \|f\|_\infty\}$$

of

$$\kappa = \sup_{f \in \mathcal{S}} \sup_{\tilde{\nu} \in \mathcal{V}} \int_X f(x)/\|f\|_\infty\,\mathrm{d}\tilde{\nu}(x)\,, \tag{3.29}$$

where $\mathcal{V} \overset{\text{def.}}{=} \{\tilde{\nu} = \frac{1}{Z} e^V\,\mathrm{d}\alpha \,:\, V \in \mathcal{S}\}$ and Z is the normalizing constant to make $\tilde{\nu}$ a probability measure. The supremum is attained by compactness of $\mathcal{S}$ and is strictly less than 1 (otherwise it should be constant α a.e. equal to 0 since $\|f\|_{\circ,\infty} = \|f\|_\infty$). □

Theorem 18 (Linear convergence of Sinkhorn)**.** *The sequence (u_k, v_k) linearly converges to (u_*, v_*) for the sup norm up to translation $\|\cdot\|_\infty^\circ$.*

Proof. The proof is a direct application of the previous property. Denote $\kappa(\alpha)$ and $\kappa(\beta)$ the contraction constants of respectively S_α and S_β, then,

$$\|S_\beta \circ S_\alpha(u_1) - S_\beta \circ S_\alpha(u_2)\|_{\circ,\infty} \le \kappa(\alpha)\kappa(\beta)\|u_1 - u_2\|_{\circ,\infty}\,, \tag{3.30}$$

therefore, the convergence is linear. □

Remark 5. *The proof of the rate of convergence implies the proof of convergence. However, it is likely that the arguments for the linear rate do not generalize in other situations such as unbalanced or multimarginal optimal transport, whereas the existence part probably adapt to such cases.*

The contraction constant κ is not explicit in Proposition 17 and we now give a quantitative estimate.

Proposition 19. *One has $\kappa(\alpha) \le 1 - e^{-\frac{2}{\varepsilon} L \operatorname{diam}(\alpha)}$, if c is L−Lipschitz and* diam(α) *is the diameter of the support of α.*

Proof. Consider $g \in \mathcal{S}$ and f a nonnegative function on X that will be detailed at the end, and define $\psi_g(t) = \int_X g \frac{e^{tf}}{\int_X e^{tf}\,\mathrm{d}\alpha}\,\mathrm{d}\alpha$. Then, by differentiation

$$\psi_g'(t) + \psi_f(t)\psi_g(t) = \psi_{fg}(t)\,, \tag{3.31}$$

and therefore

$$\psi_g(t) = \int_0^t \psi_{fg}(s) e^{-\int_s^t \psi_f(u)\,\mathrm{d}u}\,\mathrm{d}s\,. \tag{3.32}$$

Observe that, since f is nonnegative,

$$\begin{aligned} |\psi_{fg}(t)| &\leq \|g\|_\infty \int_0^t \psi_f(s) e^{-\int_s^t \psi_f(u)\,\mathrm{d}u}\,\mathrm{d}s \\ &\leq \|g\|_\infty \left(1 - e^{-\int_0^t \psi_f(u)\,\mathrm{d}u}\right) \end{aligned}$$

where the last formula is obtained by direct integration. As remarked above, the Boltzmann measure does depend on f only up to an additive constant so that we will bound $\sup f - \inf f$. Last, the term in S_α which appears in the Boltzmann measure is $\frac{1}{\varepsilon}(tu_1 + (1-t)u_2 - c(x,\cdot))$ for which a trivial bound is $\frac{2}{\varepsilon} L\,\mathrm{diam}(\alpha)$. □

Remark 6. *An explicit contraction rate in the case of measures that have a non compact support (but under other assumptions) could potentially be obtained using different norms and logarithmic Sobolev inequalities.*

3.2. **A glimpse at numerical implementation.**

4. Dynamical formulation of optimal transport

4.1. **An informal discussion on dynamic formulation.** In this section, we introduce the Benamou-Brenier formulation [Benamou and Brenier, 2000] of the Kantorovich problem. This formulation applies to distances on length spaces or more generally which can be expressed as the minimization of some Lagrangian. For instance, in the case M is a Riemannian manifold with a metric g, one can consider the induced distance squared

$$c(x,y) = \inf\left\{ \int_0^1 g_x(\dot{x},\dot{x})\,\mathrm{d}t \mid x \in C^1([0,1],M) \text{ and } (x(0),x(1)) = (x,y) \right\}\,, \tag{4.1}$$

where $\dot{x}$ denotes the time derivative of the path x. The Benamou-Brenier formulation consists in writing a similar length minimizing problem, not on the base space M, but on the space of probability measures $\mathcal{P}(M)$. We first rewrite the cost in the optimal transport functional on the space of vector fields: that is, if $\rho_1 = (\exp \varepsilon u)_*(\rho_0)$ where exp is the Riemannian exponential, that is ρ_1 is the pushforward of ρ_0 by a small perturbation of identity by a vector field u defined on M, and, assuming that the coupling is $\pi_\varepsilon = (\mathrm{id}, \mathrm{id} + \varepsilon u)_*\rho_0$, we get

$$\langle \pi_\varepsilon, d(x,y)^2 \rangle \simeq \varepsilon^2 \int_M \|v(x)\|^2\,\mathrm{d}\rho_0(x)\,. \tag{4.2}$$

Thus, one should be able to rewrite the optimal transport problem as an optimal control problem on the space of densities and where the control variable is a time dependent vector field,

$$\inf_{\rho,v} \int_0^1 \int_M \|v(t,x)\|^2\,\mathrm{d}\rho(x)\,\mathrm{d}t\,, \tag{4.3}$$

under the continuity equation constraint $\partial_t \rho(t,x) + \mathrm{div}(\rho(t,x)v(t,x)) = 0$ and time boundary constraints $\rho(0) = \rho_0$, $\rho(1) = \rho_1$. However, what is probably surprising is that we started from a convex optimization problem which we turned into a non-convex one by introducing time. Benamou and Brenier proposed a convex reformulation of the previous control problem in the following form:

$$\inf_{\rho,m} \int_0^1 \int_M \frac{\|m\|^2}{\rho}\,\mathrm{d}\rho(t,x)\,\mathrm{d}t\,, \tag{4.4}$$

under the linear constraint $\partial_t \rho(t,x) + \mathrm{div}(m) = 0$ and same time boundary constraints on ρ. The proof that the Kantorovich and Benamou-Brenier formulations are equal can be found in [Benamou and Brenier, 2000] and it is based on the convexity of the functional. This formulation was introduced by Benamou and Brenier for numerical purposes. Indeed, one can apply convex optimization algorithms to solve the formulation (4.4).

4.2. **Gradient flows.** Gradient flows with respect to the Wasserstein metric is now well-known in the litterature. We briefly present it now since it is connected with convex optimization. Maybe the most surprising fact in this section is the fact that one does not need the real Wasserstein metric (by this, we mean to refer to the Kantorovich optimization problem) in order to compute the gradient flows but instead, just the expansion in Formula (4.3). Indeed, consider a functional on the space of densities denoted by $F(\rho)$, then one may want to consider the vector field v that is acting on the current density ρ while minimizing its kinetic energy and driving F downwards. In mathematical terms,

$$\arg\min_v \frac{1}{2}\int_M \|v(t,x)\|^2 \,\mathrm{d}\rho(x) + \langle \frac{\delta F}{\delta \rho}(\rho), -\mathrm{div}(\rho v)\rangle \,, \tag{4.5}$$

where we informally denoted by $\frac{\delta F}{\delta \rho}$ the Fréchet derivative of F. Note that the previous definition generalizes the gradient for a function f defined on $\mathbb{R}^d$, $\nabla f(x) = \arg\min_w \frac{1}{2}\|w\|^2 - df_x(w)$. We get now, $v = \nabla \frac{\delta F}{\delta \rho}(\rho)$ and thus

$$\dot{\rho} = \mathrm{div}\left(\rho \nabla \frac{\delta F}{\delta \rho}(\rho)\right) \,. \tag{4.6}$$

The well-known case is the entropy $F(\rho) = \int_X \rho(x)(\log(\rho(x)) - 1)\,\mathrm{d}x$ for which $\frac{\delta F}{\delta \rho}(\rho) = \log(\rho)$ and so $\dot{\rho} = \Delta\rho$. We underline again that we only used the first order expansion of the transport cost by a velocity field in order to obtain this formal derivation of the so-called Wasserstein gradient flows. One can now define implicit gradient scheme similar to definition 8 by replacing the Hilbert norm with the Wasserstein distance, with τ a timestep parameter,

$$\rho_{k+1} = \arg\min_\rho \frac{1}{2\tau} W_2^2(\rho_k,\rho) + F(\rho) \,. \tag{4.7}$$

Remark 7. *Note again that one does not need the Wasserstein metric itself in order to get the convergence of this gradient flow to its continuous limit. Every metric on the space of densities for which the underlying metric tensor is the same than the Wasserstein distance would be suitable.*

References

[Benamou and Brenier, 2000] Benamou, J.-D. and Brenier, Y. (2000). A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393.

[Berman, 2017] Berman, R. J. (2017). The sinkhorn algorithm, parabolic optimal transport and geometric monge-amp\ere equations. *arXiv preprint arXiv:1712.03082*.

[Chizat et al., 2018] Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. (2018). An interpolating distance between optimal transport and fisher—rao metrics. *Found. Comput. Math.*, 18(1):1–44.

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

[Cuturi and Peyré, 2019] Cuturi, M. and Peyré, G. (2019). *Computational Optimal Transport*. preprint.

[Feydy et al., 2018] Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouvé, A., and Peyré, G. (2018). Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. *arXiv e-prints*, page arXiv:1810.08278.

[Papadakis et al., 2014] Papadakis, N., Peyré, G., and Oudet, E. (2014). Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238.