An elementary introduction to entropic regularization and proximal methods for numerical optimal transport

François-Xavier Vialard

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. These notes are, on purpose, written at an elementary level, with almost no prerequisite knowledge and the writing style is relatively informal. 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 with a simple proof. We prove 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. We end this short course with an overview of other dynamical formulations of optimal transport like problems.

1. INTRODUCTION

These notes are based on [Cuturi and Peyré, 2019] and most of the important references can be found there. 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 only point that differs from the usual litterature is a proof of the linear convergence of the Sinkhorn algorithm in the continuous setting, which relies on the estimation of the L^1 distance between two Gibbs measures (see Theorem 19 and Lemma 20). The last results on Sinkhorn divergence are based on [Feydy et al., 2018]. We briefly present the dynamical formulation of optimal transport, we refer to [Santambrogio, 2015] for more details. 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

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

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

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 $\to$ 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

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

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.

2.2. **Elementary convex optimization.**

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 $f_\tau(y) \stackrel{\text{def.}}{=} \min_x \frac{1}{2\tau}\|x-y\|^2 + f(x)$ 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

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

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

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

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

In particular, if it is reasonably cheap to compute the proximal operator of f, then the implicit gradient descent $x_{k+1} = \text{prox}_{\tau f}(x_k)$ 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*

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

known as Moreau's identity.

Let us be interested in the following optimization problem of a function $\mathcal{F}(x)$ that can be written as the minimization of the sum

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

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 $\langle \nabla g(x_*), x-x_*\rangle + \frac{L}{2}\|x-x_*\|^2 = \frac{L}{2}(\|x - x_* + \frac{1}{L}\nabla g(x_*)\|^2 - \frac{1}{L^2}\|\nabla g(x_*)\|^2)$,

$$x_{k+1} = \text{prox}_{(1/L)f}(x_k - \frac{1}{L}\nabla g(x_k))\,, \tag{2.21}$$

This minimization algorithm is called forward-backward, it is the composition of an explicit gradient step on g followed by an implicit gradient step of f. The convergence of this algorithm can be proven for a general step size $\tau \leq 1/L$ and the rate of convergence is in $1/k$, more precisely $\mathcal{F}(x_k) - \mathcal{F}(x_*) \leq \frac{1}{2\tau k}\|x_* - x_0\|^2$. This algorithm has an accelerated version named FISTA.

The Benamou and Brenier formula of the optimal transport problem, as described later, does not take the form of the function (2.19). In fact, it will be formulated as the minimization of the sum of two functions which are "simple". 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 central theorem in convex analysis, the Fenchel-Rockafellar theorem.

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.

2.3. **Primal-dual.** The problem of interest consists in the minimization of

$$\inf_x f(Kx) + g(x) \tag{2.26}$$

where f, g are convex, lsc and simple, which is the case we are interested in for optimal transport. In the above formulation, replace f with $(f^*)^*(x) = \max_p \langle p, Kx\rangle - f^*(p)$ to obtain

$$\inf_x \max_p \langle p, Kx\rangle - f^*(p) + g(x)\,. \tag{2.27}$$

The idea of primal-dual algorithm is to use this formulation by alternating optimization steps in x and p. More precisely, alternating an implicit step in x and an implicit step in p. For instance, the optimality condition on x reads

$$0 \in K^* p + \partial g(x) \tag{2.28}$$

which can be alternatively rewritten as

$$x - \tau K^* p \in (\mathrm{Id} + \tau \partial g)(x)\,. \tag{2.29}$$

Writing a similar equation on p leads to

$$x \leftarrow (\mathrm{Id} + \tau_1 \partial g)^{-1}(x - \tau_1 K^* p) \tag{2.30}$$

$$p \leftarrow (\mathrm{Id} + \tau_2 \partial f^*)(p + \tau_2 K x)\,, \tag{2.31}$$

where τ_1, τ_2 are the implicit gradient stepsizes. There exist different formulations and extensions of this algorithm. For instance, the primal-dual scheme

$$x_{k+1} \leftarrow \mathrm{prox}_{\tau_1 g}(x_k - \tau_1 K^* p) \tag{2.32}$$

$$p_{k+1} \leftarrow \mathrm{prox}_{\tau_2 f^*}(p_k + \tau_2 K(2x_{k+1} - x_k))\,, \tag{2.33}$$

whose convergence is guaranteed if $\tau_1 \tau_2 L^2 \leq 1$, where $\|K\| \leq L$. If more regularity on the objective function is available, acceleration of this algorithm can be used.

2.4. **Augmented Lagrangian and ADMM.** Hereafter, the objective functions are of the type

$$\min_{Ax + By = b} f(x) + g(y)\,. \tag{2.34}$$

Note that this formulation encompasses the functions of type $f(x) + g(Kx)$ via a correct choice of the linear maps A, B and the vector b. The idea of such methods is to add a Lagrange multiplier z and a quadratic penalty on the constraint with coefficient γ,

$$\min_{x,y} \sup_z f(x) + g(y) + \langle z, b - Ax - By\rangle + \frac{\gamma}{2}\|b - Ax - By\|^2\,. \tag{2.35}$$

Then, the ADMM algorithm reads

$$x_{k+1} \leftarrow \arg\min_x f(x) - \langle z_k, Ax\rangle + \frac{\gamma}{2}\|b - Ax - By_k\|^2 \tag{2.36}$$

$$y_{k+1} \leftarrow \arg\min_y g(y) - \langle z_k, By\rangle + \frac{\gamma}{2}\|b - Ax_{k+1} - By\|^2 \tag{2.37}$$

$$z_{k+1} \leftarrow z_k + \gamma(b - Ax_{k+1} - By_{k+1})\,. \tag{2.38}$$

The last step of this algorithm is a dual ascent step and its gradient is $\frac{1}{\gamma}$ Lipschitz.

2.5. **Douglas-Rachford algorithm.** This algorithm is designed for the minimization of

$$\min_x g(x) + f(x) \tag{2.39}$$

one writes

$$x_{k+1} \leftarrow \mathrm{prox}_{\tau_1 g}(x_k - \tau_1 p_k) \tag{2.40}$$

$$p_{k+1} \leftarrow \mathrm{prox}_{\tau_2 f^*}(p_k + \tau_2(2x_{k+1} - x_k))\,, \tag{2.41}$$

with $\tau_1\tau_2 \leq 1$ to ensure convergence. Then, one has, using $\tau_1\tau_2 = 1$ and Moreau's identity on $\mathrm{prox}_{\tau f^*}$,

$$x_{k+1} \leftarrow \mathrm{prox}_{\tau g}(v_k) \tag{2.42}$$

$$v_{k+1} \leftarrow v_k - x_{k+1} + \mathrm{prox}_{\tau f}(2x_{k+1} - v_k)\,. \tag{2.43}$$

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

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

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,

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

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)**.**

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

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

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

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 $\mathrm{KL}(\pi \mid \mathbf{1} \otimes \mathbf{1})$ and a simple calculation show that the argmin is independent of the choice of the measures α, β in $\mathrm{KL}(\pi \mid \alpha \otimes \beta)$. 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

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

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 (3.18)) 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 where we define the norm in oscillation of* f*,*

$$\|f\|_{\circ,\infty} \stackrel{\text{def.}}{=} \frac{1}{2}(\sup f - \inf f) = \inf_{a \in \mathbb{R}} \|f(x) - a\|_{\infty,\alpha}\,. \tag{3.22}$$

where the sup, inf and sup norm are taken w.r.t. α*. Sometimes, we use* $\operatorname{osc}(f) = (\sup f - \inf f)$.[1]

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

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

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

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

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. 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 $u_1 \neq u_2$

$$\|S_\alpha(u_1) - S_\alpha(u_2)\|_{\circ,\infty} \leq \|S_\alpha(u_1) - S_\alpha(u_2)\|_\infty < \|u_1 - u_2\|_{\circ,\infty} \tag{3.26}$$

since the case of equality implies that $u_1 = u_2$. Refining the above inequality (3.25), one has

$$|S_\alpha(u_1)(x) - S_\alpha(u_2)(x)| \leq \kappa \|u_1 - u_2\|_{\circ,\infty}\,, \tag{3.27}$$

where, κ is defined by optimization on the set

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

of

$$\kappa = \sup_{f \in \mathcal{S} \setminus \{0\}} \sup_{\tilde{\nu} \in \mathcal{V}} \frac{1}{\|f\|_\infty} \int_X f(x)\,\mathrm{d}\tilde{\nu}(x)\,, \tag{3.28}$$

where $\mathcal{V} \stackrel{\text{def.}}{=} \{\tilde{\nu} = \frac{1}{Z} e^V \mathrm{d}\alpha \ :\ V \in \frac{1}{\varepsilon}\mathcal{S}\}$ 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* $\|\cdot\|_{\circ,\infty}$*.*

[1] This notation is often used in the literature of concentration inequalities.

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,

$$\|S_\beta \circ S_\alpha(u_1) - S_\beta \circ S_\alpha(u_2)\|_{\circ,\infty} \leq \kappa(\alpha)\kappa(\beta)\|u_1 - u_2\|_{\circ,\infty}\,, \tag{3.29}$$

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 multimarginal optimal transport, whereas the existence part could adapt to such cases.*

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

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

Proof. We first give an estimation of the oscillations of $S_\alpha(f)$:

$$\frac{1}{2}\left|S_\alpha(u_1)(y) - S_\alpha(u_2)(y) - S_\alpha(u_1)(x) - S_\alpha(u_2)(x)\right| \leq \frac{1}{2}\left|\int_0^1 \langle u_1 - u_2, \nu_{t,y} - \nu_{t,x}\rangle\,\mathrm{d}t\right|\,, \tag{3.30}$$

where $\nu_{t,z} \overset{\text{def.}}{=} \frac{1}{Z}e^{\frac{t(u_1-u_2)+u_2-c(z,\cdot)}{\varepsilon}}\,\mathrm{d}\alpha$ (with Z the normalizing constant). We now use a the L^∞, L^1 bound and we note that the $\|u_1 - u_2\|_{0,\infty} \leq \|u_1 - u_2\|_\infty$. For the L^1 bound on $\nu_{t,y} - \nu_{t,x}$, we use Lemma 20. Thus , we get

$$\|S_\alpha(u_1) - S_\alpha(u_2)\|_{\circ,\infty} \leq \kappa\|u_1 - u_2\|_{\circ,\infty}\,, \tag{3.31}$$

where κ is the constant estimated in Lemma 20 below, for which the role of $u - v$ is taken by $\frac{1}{\varepsilon}(c(x,\cdot) - c(y,\cdot))$ and a trivial bound is $\|u - v\|_{\circ,\infty} \leq \frac{1}{\varepsilon}L\,\mathrm{diam}(\alpha)$. □

Lemma 20. *Let u, v be two continuous functions on X and α be a probability measure and denote ν_u, ν_v the Boltzmann measures associated with u, v, which are $\nu_u = \frac{1}{Z_u}e^u\alpha$ and $\nu_v = \frac{1}{Z_v}e^v\alpha$ then*

$$\|\nu_u - \nu_v\|_{L^1} \leq 2(1 - e^{-\|u-v\|_{\circ,\infty}}) = 2(1 - e^{-\frac{1}{2}\,\mathrm{osc}(u-v)})\,. \tag{3.32}$$

Proof. Consider g a bounded function on X and define $\psi_g(t) = \int_X g\frac{e^{tv+(1-t)u}}{\int_X e^{tv+(1-t)u}\,\mathrm{d}\alpha}\,\mathrm{d}\alpha$. Then, by differentiation

$$\psi_g'(t) + \psi_{v-u}(t)\psi_g(t) = \psi_{(v-u)g}(t)\,, \tag{3.33}$$

and therefore

$$e^{\int_0^t \psi_{v-u}(s)\,\mathrm{d}s}\psi_g(t) - \psi_g(0) = \int_0^t \psi_{(v-u)g}(s)e^{\int_0^s \psi_{v-u}(r)\,\mathrm{d}r}\,\mathrm{d}s\,. \tag{3.34}$$

Observe that, since one can assume (the Boltzmann measures are defined up to an additive constant on the function) that $u - v$ is nonnegative,

$$\begin{aligned}|e^{\int_0^t \psi_{u-v}(s)\,\mathrm{d}s}\psi_g(t) - \psi_g(0)| &\leq \|g\|_\infty \int_0^t \psi_{u-v}(s)e^{\int_0^s \psi_{u-v}(r)\,\mathrm{d}r}\,\mathrm{d}s \\ &\leq \|g\|_\infty\left(e^{\int_0^t \psi_{u-v}(s)\,\mathrm{d}s} - 1\right)\end{aligned}$$

where the last formula is obtained by direct integration. Now, by exchanging the role of u, v, only two cases are possible: Whether $\psi_g(1) \geq \psi_g(0) \geq 0$ or $\psi_g(1) \geq 0 \geq \psi_g(0)$. In the first case, one has

$$|e^{\int_0^t \psi_{u-v}(s)\,\mathrm{d}s}(\psi_g(t) - \psi_g(0))| \leq |e^{\int_0^t \psi_{u-v}(s)\,\mathrm{d}s}\psi_g(t) - \psi_g(0)| \leq \|g\|_\infty\left(e^{\int_0^t \psi_{u-v}(s)\,\mathrm{d}s} - 1\right)\,. \tag{3.35}$$

In the second case, there exists $t_0 \in [0,1]$ such that $\psi_g(t_0) = 0$, and thus

$$|\psi_g(1)| \leq \|g\|_\infty \left(1 - e^{-\int_{t_0}^1 \psi_{u-v}(s)\,\mathrm{d}s}\right)$$
$$|\psi_g(0)| \leq \|g\|_\infty \left(1 - e^{-\int_0^{t_0} \psi_{u-v}(s)\,\mathrm{d}s}\right)$$

and therefore, by optimizing[2] on the parameter t_0, we obtain

$$|\psi_g(1) - \psi_g(0)| \leq |\psi_g(1)| + |\psi_g(0)| \leq 2\|g\|_\infty (1 - e^{-\frac{1}{2}\int_0^1 \psi_{u-v}(s)\,\mathrm{d}s}) . \tag{3.36}$$

Since $\psi_{u-v}(t) \leq 2\|u-v\|_{\circ,\infty}$, we get, in the two cases

$$\|\nu_u - \nu_v\|_{L^1} \leq 2(1 - e^{-\|u-v\|_{\circ,\infty}}) . \tag{3.37}$$

□

Remark 6. *In fact, the bound on $\psi_{u-v}(t)$ is not sharp since, here again, $|\langle (u-v), \nu_{u-v}\rangle| < \|u-v\|_\infty$ unless $u - v = cste$. In this case, this would imply that the cost is a constant function which is not an interesting case to consider. Indeed, the optimal coupling is the product of marginals.*

3.2. **Hilbert metric and convergence in the discrete setting.** In this paragraph, we give a brief description of the usual proof of convergence of the contraction rate in a discrete setting.

Definition 11 (Hilbert metric)**.** Let $\mathbb{R}^n_{++}$ be the cone of positive coordinates vector. The Hilbert metric on this cone is

$$\mu(x,y) \stackrel{\text{def.}}{=} \max_{i,j} \log\left(\frac{x_i y_j}{x_j y_i}\right) . \tag{3.38}$$

A few remarks are in order: the quantity μ is nonnegative since one can take $i = j$ in Formula (3.38) to get $\mu(x,y) \geq \log(1) = 0$ and $\mu(x,\lambda x) = 0$, therefore the Hilbert metric cannot be a metric on $\mathbb{R}^n_{++}$ but rather, it is a metric on $\mathbb{R}^n_{++}/\mathbb{R}_{>0}$, i.e. quotienting by multiplication by positive scalars. Thus, it is said to be a projective metric, a metric on the space of lines, or more precisely in this case, half-lines. Remark that if $\mu(x,y) = 0$ then it implies that $\forall i,j$ one has $\frac{x_i}{y_i} = \frac{x_j}{y_j}$ therefore, this quantity being independent of the index, one has $x = \lambda y$ for a positive real λ. Last, the triangle inequality is simple to obtain and ensures that the Hilbert metric indeed is a metric on $[\mathcal{S}_n]_{++} \stackrel{\text{def.}}{=} \mathcal{S}_n \cap \mathbb{R}^n_{++}$, which is one possible parametrization of this quotient space. An important fact concerning the Hilbert metric is the following:

Theorem 21. *The set $[\mathcal{S}_n]_{++}$ endowed with the Hilbert metric is complete.*

Proof. We refer the reader to [Nussbaum, 1987]. □

Obviously, this theorem is non trivial since $[\mathcal{S}_n]_{++}$ is an open set of $\mathbb{R}^n$. This fact is a key ingredient of the celebrated Birkhoff theorem:

Theorem 22. *Let $A \in \mathbb{R}^{m\times n}_{++}$ be a matrix with positive coefficients, then*

$$\mu(Ax, Ay) \leq \kappa(A)\mu(x,y) \forall x, y \in \mathbb{R}^n_{++} \tag{3.39}$$

where the constant $\kappa(A) = \tanh(\frac{\Delta(A)}{4}) < 1$ and

$$\Delta(A) = \max_{i,j} \mu(Ae_i, Ae_j) = \max_{ijkl} \log\left(\frac{A_{ik}A_{jl}}{A_{il}A_{jk}}\right) . \tag{3.40}$$

The constant $\kappa(A)$ can be alternatively written as $\kappa(A) = \frac{e^{\Delta(A)/2}-1}{e^{\Delta(A)/2}+1}$. The Perron-Frobenius theorem is a corollary of Birkhoff's theorem:

[2]Optimality is attained when the two quantities in the exponential are equal, that is $\int_{t_0}^1 \psi_{u-v}(s)\,\mathrm{d}s = \int_0^{t_0} \psi_{u-v}(s)\,\mathrm{d}s = \frac{1}{2}\int_0^1 \psi_{u-v}(s)\,\mathrm{d}s$.

Theorem 23. *Let $A \in \mathbb{R}^{n\times n}_{++}$ be a square matrix with positive coefficients and $x_0 \in \mathbb{R}^n_{++}$. The sequence $x_{k+1} = \frac{Ax_k}{\|Ax_k\|}$ converges linearly to the unique solution which is an eigenvector associated with the spectral radius eigenvalue of A. In particular, $\mu(x_k, x_*) \leq c\kappa(A)^k$.*

The important consequence of Birkhoff theorem is the linear convergence of Sinkhorn since the Gibbs kernel matrix is $k = e^{-C_{ij}/\varepsilon}$ which has positive entries. In order to see this, we insists on the following properties of the Hilbert metric:

Proposition 24. *Pointwise multiplication on $\mathbb{R}^n_{++}$ (that is $(x \cdot y)_i = x_i y_i$) as well as inversion ($(x^{-1})_i = 1/x_i$) are isometries for the Hilbert metric.*

Proof. The proof consists in a direct check of the formula (3.38). □

Let us sketch the use of these two properties to get the linear convergence for the discrete Sinkhorn algorithm.

Theorem 25. *The discrete Sinkhorn algorithm* (3.9) *linearly converges to its unique solution.*

Proof. Consider the sequences D_1^k and D_2^k generated by the Sinkhorn algorithm (3.9). One has

(3.41)
$$\mu(D_2^k, D_2^{k+1}) = \mu(\mathbf{1}_m./(A^T D_1^k), \mathbf{1}_m./(A^T D_1^{k+1})) = \mu(A^T D_1^k, A^T D_1^{k+1}) \leq \kappa(A^T)\mu(D_1^k, D_1^{k+1}).$$

Therefore, iterating this argument leads to

$$\mu(D_2^k, D_2^{k+1}) \leq \kappa(A)^2 \mu(D_2^k, D_2^{k-1}) \tag{3.42}$$

where we used the fact that $\kappa(A^T) = \kappa(A)$. The rest of the proof follows from standard arguments on contractions. □

In practice, the quantity $\kappa(A)$ can be quantified for the Sinkhorn algorithm as follows, if c is a cost which is L Lipschitz on the domain with bounded diameter D, after a Taylor expansion when $\frac{2}{\varepsilon}LD >> 1$,

$$\Delta(A) \leq \frac{2}{\varepsilon} LD \text{ and } \kappa(A) \simeq (1 - e^{-\frac{1}{\varepsilon}LD})^2 \simeq 1 - 2e^{-\frac{1}{\varepsilon}LD}. \tag{3.43}$$

It can be compared with the constant we get in Proposition 19, $\kappa = 1 - e^{-\frac{1}{\varepsilon}L\,\mathrm{diam}(\alpha)}$. The constant obtained by the Birkhoff theorem is slightly better than the one obtained by our simple computation. The latter could probably be refined to match the one given by Birkhoff's theorem by improving the bound on the entropy term $\psi_f(t)$ in the proof of Lemma 20. Indeed, the bound we gave rely on the inequality $\psi_f(t) \leq \|f\|_\infty$, but, here again, the inequality might be strict in some cases, whence the potential gain.

3.3. **A glimpse at numerical implementation.** There are different applications of the Sinkhorn regularized optimal transport: in some cases, such as machine learning, the smoothness property is an important feature and due to sometimes high-dimensional data, medium/large epsilon are useful in practice. In such a case, the matrix-vector multiplication algorithm (3.9), which has of course a computational cost less than $O(N^2)$, is appealing since it is GPU friendly and highly parallelizable.

(1) **Measures on a grid:** When the cost is separable, for instance, $c(x, y) = \sum_{i=1}^d |x_i - y_i|^2$ on $\mathbb{R}^d$, the computational complexity can be reduced. For example, in dimension 2, if one has a vector of size $N = N_1 N_2$, one can first reshape the vector in a matrix of size $(N1, N2)$, convolve with the gaussian kernel $e^{-|x_1-y_1|^2/\varepsilon}$ in the first dimension, which has the cost lower than $N_1^2 N_2$. Applying this in larger dimension d leads to a computational cost lower than $O(N^{1+1/d})$ instead of $O(N^2)$, for naive implementation.
(2) **Large cloud of points:** This situation (typically 10^5 points) differs from the previous one since the separability trick cannot be applied since the points are not on a mesh. A feasible solution consists in recomputing the kernel in the log-sum-exp computations (see below). It

has been implemented in the pytorch package GeomLoss and KeOps [Charlier et al., 2018]. We highly recommend the reader to visit this webpage.

In theory, the rate of convergence of the Sinkhorn algorithm degrades when ε is small, it is also observed in practice. For small ε, the computation needs to be done in Log-Sum-Exp formulation as in the proof of convergence to avoid overflow issues. Indeed, the iterates stay bounded, essentially due to the 1-Lipschitz property. The drawback of this formulation is that the matrix-vector multiplication algorithm (3.9) is not available any longer and as a consequence, one cannot use optimized and parallelized implementations of matrix multiplication.

4. Dynamical formulation of optimal transport

In this section, we discuss formulations of optimal transport and related evolution flows (gradient flows) that involves a time variable. For a more mathematical and complete presentation, we refer to [Santambrogio, 2015].

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

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

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).

Let us discuss informally yet another way to obtain the dynamic formulation. The Kantorovich formulation is the relaxation of the Monge formulation which can be stated as, for ρ_0, ρ_1 with density w.r.t. the Lebesgue measure

$$W_2^2(\rho_0,\rho_1) = \inf_{\varphi} \left\{ \|\varphi - \mathrm{Id}\,\|^2_{L^2(\rho_0)}\,;\ \varphi_*\rho_0 = \rho_1 \right\} \tag{4.5}$$

Using the length space property of L^2 (note that this property is based on the length space property of the distance)

(4.6)
$$\|\varphi - \mathrm{Id}\|^2_{L^2(\rho_0)} = \inf_{\varphi(t)} \int_0^1 \|\partial_t \varphi\|^2_{L^2(\rho_0)} \,\mathrm{d}t = \inf_{\varphi} \int_0^1 \int_M |\partial_t \varphi \circ \varphi^{-1}(t,y)|^2 \,\mathrm{Jac}(\varphi^{-1}) \rho_0 \circ \varphi^{-1}(y) \,\mathrm{d}y \,\mathrm{d}t\,.$$

where $\varphi(t)$ is a path between Id and φ. The term $\mathrm{Jac}(\varphi^{-1})\rho_0 \circ \varphi^{-1}(y)$ is the advected density and $\partial_t \varphi \circ \varphi^{-1}(t,y) = v(t,y)$ is the velocity field. Therefore, one obtains Formulation (4.3). Obviously, this formulation only involves kinetic energy; it is obviously possible to introduce a potential energy on the space of densities, such as the Fisher information $V(\rho) = \int_M |\nabla(\log \rho)|^2 \,\mathrm{d}\rho(x)$.

4.2. **Gradient flows.** Gradient flows with respect to the Wasserstein metric is now a well-known and well studied subject. We briefly present it now from an unformal point of view 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,

$$\arg\min_v \frac{1}{2} \int_M \|v(t,x)\|^2 \,\mathrm{d}\rho(x) + \left\langle \frac{\delta F}{\delta \rho}(\rho), -\mathrm{div}(\rho v) \right\rangle\,, \tag{4.7}$$

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

$$\dot{\rho} = \mathrm{div}\left(\rho \nabla \frac{\delta F}{\delta \rho}(\rho)\right)\,. \tag{4.8}$$

The well-known case is the entropy+potential $F(\rho) = \int_X \rho(x)(\log(\rho(x)) - 1)\,\mathrm{d}x + \int_X V(x)\rho(x)\,\mathrm{d}x$ for which $\frac{\delta F}{\delta \rho}(\rho) = \log(\rho) + V(x)$ and so

$$\dot{\rho} = \Delta \rho + \mathrm{div}(\rho \nabla V)\,,$$

which is the Fokker-Planck equation. 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,

$$\rho_{k+1} = \arg\min_\rho \frac{1}{2\tau} W_2^2(\rho_k, \rho) + F(\rho)\,. \tag{4.9}$$

The convergence of this time discrete scheme in the case of entropy has been proven by Jordan, Kinderlehrer and Otto.

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.*

Remark 8. *One particular interest of such a variational formulation is that it is possible to model evolution equations for which the corresponding PDE is somewhat singular.*

4.3. **A proximal algorithm for the dynamical formulation.** One way to numerically solve the dynamical formulation of optimal transport consists in formulating a discrete functional approximating the continuous setting, on which convex optimization algorithms can be applied. The continuous formulation can be written as

$$W_2(\rho_0, \rho_1)^2 = \inf_{\rho, m} K(\rho, m) + \iota_C(\rho, m)\,. \tag{4.10}$$

FIGURE 1. The red crosses stand for the centered grid while the blue dots are for the staggered grid

where C is the convex set of ρ, m that are time dependent quantities such that $\partial_t \rho + \mathrm{div}(m) = 0$ and $\rho(0) = \rho_0$ and $\rho(1) = \rho_1$. The quantity $K(\rho, m)$ represents the kinetic energy $\frac{1}{2}\int_0^1 \int_M \frac{\|m\|^2}{\rho}\, \mathrm{d}\rho(t,x)\, \mathrm{d}t$. In computational fluid dynamics, the method of staggered grid is often used for discretizing the continuity equation. This method makes use of two different grids for discretization: the centered grid and the staggered grid, see Figure 4.3. When the size of the problem is not too large, this is the method of choice for solving Poisson equation. We are going to discretize the equations using finite differences[3]. Let us assume that we have a quantity s defined on the staggered grid, that is $s(i+1/2)$ for $i \in [-1, n]$ for a 1D centered grid defined on $[0, n]$. Then, the divergence operator applied to s will map the staggered quantity on the centered grid:

$$\begin{aligned} \mathrm{div} : \text{Staggered} &\to \text{Centered} \\ s &\mapsto s(i+1/2+1) - s(i+1/2)\,. \end{aligned}$$

The discrete adjoint div^* is thus defined as

$$\begin{aligned} \mathrm{div}^* : \text{Staggered} &\to \text{Centered} \\ c &\mapsto -[c, 0] + [0, c]\,, \end{aligned}$$

where the notation $[0, c]$ indicates the concatenation of 1D tensors $[0]$ and c. Then, the constraint $\partial_t \rho + \mathrm{div}(m) = 0$ can be rewritten as $\mathrm{div}_{t,x}(\rho, m) = 0$ and, for each direction (time and space), there is a corresponding staggered grid: ρ is staggered in time and m is staggered in space.

Then, we have left the question how to switch between the two representations of the data: staggered and centered. We simply use the interpolation operator to go from staggered to centered grid representation:

$$\begin{aligned} \mathcal{I} : \text{Staggered} &\to \text{Centered} \\ s &\mapsto \frac{1}{2}(s(i+1/2) + s(i-1/2))\,. \end{aligned}$$

Then, one can propose the following form of the functional, denoting ρ, m the unknowns and $\tilde{\rho}, \tilde{m}$ their staggered versions,

$$\min_{(\rho, m, \tilde{\rho}, \tilde{m})} K(\rho, m) + \iota_C(\tilde{\rho}, \tilde{m}) + \iota_{\mathrm{interp}}((\rho, m), (\tilde{\rho}, \tilde{m})) \tag{4.11}$$

where ι_C is the convex indicator function of the set

$$\{(\tilde{\rho}, \tilde{m}) \mid \mathrm{div}(\tilde{\rho}, \tilde{m}) = 0 \text{ and } \tilde{\rho}(:, -1/2) = \rho_0 \text{ and } \tilde{\rho}(:, N - 1/2) = \rho_1\}\,.$$

and the function ι_{interp} is the convex indicator of the set $\{((\tilde{\rho}, \tilde{m}), (\rho, m)) \mid \mathcal{I}(\tilde{\rho}, \tilde{m}) = (\rho, m)\}$. Now, the goal is to apply convex optimization algorithms to the functional (4.11). Note that K is not a smooth convex function, and the two other functions are convex indicators. These functions are fortunately simple, in the sense that the proximal operator can be computed relatively easily. In particular, one can use the decomposition $G_1 = K + \iota_C$ and $G_2 = \iota_{\mathrm{interp}}$. In order to apply first order algorithms, we need to compute the proximal operators associated with G_1 and G_2.

In general, $\mathrm{prox}(\iota_C) = p_C$ the orthogonal projection on C. Let us detail the case of $C = \{(x, y) \mid y = Ax\}$ which is the case of ι_{interp}. Let us compute

$$\min_x \frac{1}{2}\|x - x_0\|^2 + \frac{1}{2}\|Ax - y_0\|^2\,. \tag{4.12}$$

Optimality implies

$$x - x_0 + A^*(Ax - y_0) = 0\,, \tag{4.13}$$

[3]More involved discretization could be envisaged at this point.

and thus

$$x = (\mathrm{Id} + A^*A)^{-1}(A^*y_0 + x_0)\,. \tag{4.14}$$

It is possible to use LU factorization and separability in the case of the interpolation map to speed up the computations.

The second projection we have to compute is the one associated with ι_C. One can write

$$A(\tilde\rho, \tilde m) = \begin{pmatrix} \mathrm{div}(\tilde\rho, \tilde m) \\ s_{BC}(\tilde\rho, \tilde m) \end{pmatrix} = \begin{pmatrix} 0 \\ b_0 \end{pmatrix}\,, \tag{4.15}$$

where s_{BC} stands for the evaluation of the boundary values. Therefore,

$$\mathrm{prox}_{\iota_C}(z) = \arg\min_x \frac{1}{2}|x - z|^2 \text{ s.t. } Ax = \begin{pmatrix} 0 \\ b_0 \end{pmatrix}\,. \tag{4.16}$$

Using Lagrange multipliers, the optimality condition leads to

$$x = z + A^*p \tag{4.17}$$

$$Ax = Az + AA^*p = \begin{pmatrix} 0 \\ b_0 \end{pmatrix}\,. \tag{4.18}$$

which implies

$$AA^*p = \begin{pmatrix} 0 \\ b_0 \end{pmatrix} - Az\,. \tag{4.19}$$

Remark 9. *A priori, AA^* is not invertible since $A^* : \mathbb{R}^N \mapsto \mathbb{R}^M$ with $N > M$. However, it is a symmetric nonnegative matrix and it has a pseudo-inverse.*

Indeed, AA^* is invertible on $(\mathrm{Ker}(A^*))^\perp = \mathrm{Im}(A)$ and $v_0 \in \mathrm{Im}(A)$ implies $p = (AA^*)^{-1}(v_0 - Az)$ is uniquely defined. Then,

$$x = z + A^*(AA^*)^{-1}(v_0 - Az)\,. \tag{4.20}$$

For this concrete application, we parameterize $x = x_0 + b_0$ and we use the notation $p_{\overline{BC}}(x) = x$ outside the boundaries and 0 on the boundaries. Then, with $A = \mathrm{div} \circ p_{\overline{BC}}$ we have

$$\frac{1}{2}|x - p_{\overline{BC}}(x_0)|^2 + \langle p, Ax - Ab_0\rangle \tag{4.21}$$

and get

$$\mathrm{div}\, p_{\overline{BC}} p^*_{\overline{BC}} \mathrm{div}^*\, p = Ab_0 - Ax_0\,, \tag{4.22}$$

which is a Poisson equation with Neumann boundary conditions.

We now compute the proximal operator of the kinetic energy $\sum_{\text{centered grid}} \frac{1}{2}\frac{|m|^2}{\rho}$. The first remark is that the proximal operator is applied pointwise on the grid since this is a direct sum and it amounts to computing the proximal operator of a 1D function. Just for sake of completeness, we perform the computation below

$$\arg\min_{\rho,m} \frac{1}{2\tau}|m_0 - m|^2 + \frac{1}{2\tau}|\rho - \rho_0|^2 + \frac{1}{2}\frac{|m|^2}{\rho}\,. \tag{4.23}$$

Variations in m and ρ lead to

$$\frac{1}{2\tau}(m - m_0) + \frac{m}{\rho} = 0 \tag{4.24}$$

$$\frac{1}{2\tau}(\rho - \rho_0) + \frac{1}{2}\frac{|m|^2}{\rho^2} = 0\,. \tag{4.25}$$

These two equations imply the two following relations

$$m = \frac{m_0}{(1 + \frac{2\tau}{\rho})} \tag{4.26}$$

and

$$(\rho + 2\tau)^2(\rho - \rho_0) - \tau\rho^2|m_0|^2 = 0\,. \tag{4.27}$$

By uniqueness of the proximal map, the argmin is the unique (if it exists) positive root of Equation (4.27). Otherwise, the proximal is $(\rho, m) = (0, 0)$. The computation of this 3rd order polynomial root is given in close form and it has to be done pointwise on the grid.

Remark 10. *In fact,* $\frac{|m|^2}{\rho}$ *being one-homogeneous, the Legendre-Fenchel conjugate is the convex indicator*

$$C_0 := \{(\alpha, \beta) \,|\, \alpha + \frac{1}{2}|\beta|^2 \leq 0\}\,. \tag{4.28}$$

Using these proximal maps, one can use primal-dual, Douglas-Rachford algorithms to solve the problem.

4.4. **Other dynamical formulations.** Following the Benamou and Brenier formulation, there has been lots of models proposed in the litterature deriving from it. We give hereafter two examples of such models. The first

4.4.1. *Unbalanced optimal transport.* We choose to present the extension of optimal transport to unbalanced optimal transport, that is optimal transport with creation/delation of mass. Another formulation of the problem is "how to define an extension of optimal transport for marginals that do not have the same total mass?". A possible way to go is to relax the marginal constraints in the static formulation using a divergence such as Kullback-Leibler. It is particularly nice for numerics and for the extension of the Sinkhorn algorithm. However, the difficulty is, for instance, to prove that the resulting object leads to a distance on the space of positive Radon measures. Another way to go would be to start from the Benamou-Brenier formulation which is of particular interest since it gives access to the Riemannian like metric tensor of optimal transport. Then, modify the Riemannian tensor in order to give the possibility of creation/destruction of mass. Namely, the creation/destruction of mass can be introduced via the continuity equation

$$\partial_t\rho + \operatorname{div}(\rho v) = \alpha\rho \tag{4.29}$$

where we introduced a source term parametrized by the growth rate α which depends both on time and space. Then, we have to postulate[4] a Lagrangian on this growth rate and a natural action for this is the Fisher-Rao functional

$$\frac{1}{2}\int_M \alpha^2 \,\mathrm{d}\rho\,. \tag{4.30}$$

With this Lagrangian, the extension of the Benamou-Brenier formulation is as follows, minimize, under Equation (4.29), the action

$$\inf_{\rho,v,\alpha} \int_0^1 \int_M \frac{1}{2}(\|v(t,x)\|^2 + \frac{1}{4}\alpha(t,x)^2)\,\mathrm{d}\rho(t,x)\,\mathrm{d}t\,, \tag{4.31}$$

where we emphasize the dependence of the control variable on time and space. Interestingly, this optimization problem is a slight modification of the Benamou-Brenier formulation and the same numerical framework can be used to solve the problem. Something that is not clear from this dynamic formulation is the existence of a Kantorovich formulation of the problem. A possible way to realize that it is the case is to start from the ansatz that ρ is a Dirac mass for all time: $\rho = m(t)\delta_{x(t)}$ then the Lagrangian reduces to $m\,\mathrm{d}x^2 + \frac{1}{4}\frac{\mathrm{d}m^2}{m}$, which can be transformed into $r^2\,\mathrm{d}x^2 + \mathrm{d}r^2$ with the change of variable $m = r^2$. This metric is a polar coordinate metric for which the change of variables re^{ix} can be used. Therefore the distance is explicit $d^2(r_0^2\delta_{x_0}, r_1^2\delta_{x_1}) = \left|r_0e^{ix_0} - r_1e^{ix_1}\right|^2$. Using convexity and 1-homogeneity of the functional, it can be proven that there exists a Kantorovich formulation associated with the dynamic formulation defined above. Given ρ_1, ρ_2 two positive

[4]Other Lagrangian can be postulate but to make it well-defined on the space of measures, it is important to have a one-homogeneous functional.

Radon measures, the associated quantity is a distance on the space of positive Radon measures and is given by the Wasserstein-Fisher-Rao metric (also known as Hellinger-Kantorovich),

$$WFR(\rho_1,\rho_2)^2 = \inf_{\pi} \mathrm{KL}(\pi_1,\rho_1) + \mathrm{KL}(\pi_2,\rho_2) + \langle \pi, -\log(\cos^2(\min(d(x,y), \frac{\pi}{2}))\rangle\,, \tag{4.32}$$

where the optimization is performed on π which is a positive Radon measure on the product space $M \times M$. For the proof of this theorem, we refer the reader to [Chizat et al., 2015]. The surprising fact in the Kantorovich formulation above is the cost which appears in the scalar product[5]. One can replace it with the squared distance while still preserving the metric property of the resulting quantity. However, the length space implied by this metric known as Gaussian-Hellinger is the one given by WFR (4.32), therefore it shows the importance of the WFR formulation (this fact is proven in [Liero et al., 2015]). Obviously, this formulation is amenable to entropic regularization with associated Sinkhorn algorithms for which linear convergence can be proven also for more general divergence terms.

4.4.2. *Entropic regularization and generalized incompressible Euler flows.* Interestingly, the entropic regularization has also a dynamic formulation on the space of densities. One has the equality

$$\mathrm{OT}_\varepsilon(\rho_0,\rho_1) = \inf_{\rho,v} \int_0^1 \int_M \left(\frac{1}{2}|v|^2 + \frac{\varepsilon}{2}|\nabla \log(\rho)|^2\right) \mathrm{d}\rho\,\mathrm{d}t\,, \tag{4.33}$$

under the continuity equation constraint $\partial_t\rho + \mathrm{div}(\rho v) = 0$. The term in $\nabla\log(\rho)$ is a potential term (in contrast to the kinetic energy term), it is known as the Fisher information. Optimality is attained for a vector field v which is a gradient field and one has the following system

$$\begin{cases} \partial_t\rho + \mathrm{div}(\rho\nabla p) = 0 \\ \partial_t p + \frac{1}{2}|\nabla p|^2 = \frac{\delta}{\delta\rho}\left(\frac{\varepsilon}{2}\int_M |\nabla\log(\rho)|^2)\,\mathrm{d}\rho\right)\,. \end{cases} \tag{4.34}$$

Interestingly, this system can be transformed by introducing the following change of variables $z = p - \log(\rho)$[6] to get

$$\begin{cases} \partial_t\rho + \mathrm{div}(\rho\nabla p) = \Delta\rho \\ \partial_t z + \frac{1}{2}|\nabla z|^2 = -\Delta z\,. \end{cases} \tag{4.35}$$

The reader could be surprised of the minus sign in the second equation, however, this equation is to be understood as an adjoint equation which is read backward in time. Recent numerical algorithms have been proposed to solve the formulation (4.33) which is smooth and strongly convex on some bounded sets (depending on the initial and final conditions) due to the entropic term. In particular acceleration methods in convex optimization can be used.

5. Further developments around entropy regularized OT

This discussion is based on [Feydy et al., 2018] in which we study new divergences on the space of probability for applications to machine learning. The motivation is to use the computational efficiency of Sinkhorn algorithm while still retaining important mathematical properties: In particular, the Wasserstein L^2 distance metrizes the weak-* convergence on the space of probability measures on a compact metric space. Recall that, on a compact metric space, the weak-* convergence of μ_n to μ is written $\mu_n \rightharpoonup \mu$ and is defined by duality with continuous functions $C(X)$, $\langle f,\mu_n\rangle \to \langle f,\mu\rangle$ for every $f \in C(X)$. Convergence in L^2 Wasserstein distance is equivalent to weak-* convergence. Recall that the L^1 Wasserstein distance has a dual formulation on the space of 1-Lipschitz functions f, $W_1(\mu,\nu) = \sup\{\langle f,\mu-\nu\rangle; \mathrm{Lip}(f) \le 1\}$. If instead of maximizing this quantity over f in the 1-Lipschitz ball, one instead chooses $f \in B_H$, with H a Reproducing Kernel Hilbert Space (RKHS) such as Sobolev spaces (of sufficiently high degree of smoothness), one obtains Maximum

[5]To give the rough idea of where this cost comes from, there exists a corresponding optimal transport problem on the space of diracs in mass and position (i.e. (m,x)) that involves functions that are one-homogeneous in the radial direction. It then leads to an inequality on quadratic functions which implies that a particular discriminant is nonnegative.

[6]Sometimes, the quantity $\nabla\log(\rho)$ is called the osmotic velocity, see for instance Nelson's book [Nelson, 1967].

Mean Discrepancies (MMD), well-known in the Machine Learning litterature, which also metrizes the convergence in law. Although this is a common feature between MMD and OT, there are two important differences, for instance in the discrete setting,

(1) MMD distances are smooth with respect to the position of Dirac masses which is not the case for OT.
(2) With respect to the position of the Dirac masses, OT has more convexity properties than MMD (indeed, if the two input measures differ from a translation (which is the optimal map), then the OT cost is convex with respect to the translation).

The smoothness property is important for the use of smooth optimization methods and in particular the use of automatic differentiation. Then, convexity is important for convergence towards a global optimum when doing gradient descent with respect to the position of Dirac masses. It is possible to define new divergences based on entropy regularized optimal transport that interpolates between OT and MMD. We refer to [Feydy et al., 2018] for more background and motivations and we only state the main result.

Theorem 26. *Define*

$$S_\varepsilon(\mu,\nu) = OT_\varepsilon(\mu,\nu) - \frac{1}{2}(OT_\varepsilon(\mu,\mu) + OT_\varepsilon(\nu,\nu))\,. \tag{5.1}$$

If the cost c in definition of OT_ε defines via $e^{-\frac{1}{\varepsilon}c(x,y)}$ a positive universal kernel then S_ε is a symmetric positive definite loss function which is smooth with respect to both input measures, as well as convex with respect to each of the inputs (i.e. coordinatewise).

Due to the use of the Sinkhorn algorithm to compute each term in the definition of S_ε, it makes this new divergence a computable smooth approximation of optimal transport. For more details on the actual algorithm, we refer to [Feydy et al., 2018]. Importantly, the gradient has a closed form and is defined in the continuous setting. In particular, automatic differentiation can be overriden if needed, however, its accuracy depends on the convergence of the Sinkhorn algorithm.

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*.

[Charlier et al., 2018] Charlier, B., Feydy, J., and Glaunes, J. (2018). Kernel operations on the gpu, with autodiff, without memory overflows.

[Chizat et al., 2015] Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. (2015). Unbalanced optimal transport: geometry and kantorovich formulation. *arXiv preprint arXiv:1508.05216*.

[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.

[Liero et al., 2015] Liero, M., Mielke, A., and Savaré, G. (2015). Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, pages 1–149.

[Nelson, 1967] Nelson, E. (1967). *Dynamical theory of Brownian motion*. Princeton University Press.

[Nussbaum, 1987] Nussbaum, R. D. (1987). Iterated nonlinear maps and hilbert's projective metric: A summary. In Chow, S.-N. and Hale, J. K., editors, *Dynamics of Infinite Dimensional Systems*, pages 231–248, Berlin, Heidelberg. Springer Berlin Heidelberg.

[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.

[Santambrogio, 2015] Santambrogio, F. (2015). *Optimal Transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their applications*. Springer.