ELEMENTARY INTRODUCTION TO SOME TOPICS IN NUMERICAL OPTIMAL TRANSPORT

FRANÇOIS-XAVIER VIALARD

ABSTRACT. This note contains the material I presented at the summer school on optimal transport, TU Dortmund in 2023 and what I presented to the CEA-EDF-INRIA summer school about numerical optimal transport in 2019. It is, on purpose, written at an elementary level, with almost no prerequisite knowledge in optimal transport and the writing style is informal. All the methods presented hereafter rely on convex optimization, so we give a basic introduction to convex analysis and optimization. We discuss gradient flows in Wasserstein space after an introduction to the (infinite dimensional) geometry of the Wasserstein space and the Benamou-Brenier formulation. 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.

1. INTRODUCTION

These notes[1] 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 7 and Lemma 8). 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]. We present the different formulations of unbalanced optimal transport, static and dynamic and conic in the particular case of the relative entropy as a penalty for the marginal constraints (Wasserstein-Fisher-Rao and generalizations). We also present the dynamic formulation of entropic regularization of OT and we present a corresponding formulation in the case of the Wasserstein-Fisher-Rao metric.

2. 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

$$\mathcal{S}_{n,m} = \{\pi_{ij} \in \mathbb{R}_+^{n\times m} \,:\, \sum_{i=1}^{n}\sum_{j=1}^{m} \pi_{ij} = 1\}\,. \tag{2.1}$$

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

$$\mathrm{OT}(\rho_1,\rho_2) = \min\langle \pi(i,j), c(i,j)\rangle \text{ such that } \sum_j \pi_{i,j} = \rho_1(i) \text{ and } \sum_i \pi_{i,j} = \rho_2(j) \quad \forall i,j\,. \tag{2.2}$$

[1] I am thankful to Théo Dumont, who improved substantially the writing of this note.

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

$$\min_{\pi \in \Pi(\rho_1, \rho_2)} \langle \pi(i,j), c(i,j) \rangle - \varepsilon \, \mathrm{Ent}(\pi) \,, \tag{2.3}$$

where we denoted the set of admissible couplings by

$$\Pi(\rho_1, \rho_2) \stackrel{\text{def.}}{=} \{ \pi \in \mathcal{S}_{n,m} \, : \, \sum_j \pi_{i,j} = \rho_1(i) \text{ and } \sum_i \pi_{i,j} = \rho_2(j) \forall i, j \} \,. \tag{2.4}$$

and the Shannon entropy, which is a strictly concave function

$$\mathrm{Ent}(\pi) \stackrel{\text{def.}}{=} - \sum_{i,j} \pi_{i,j} (\log(\pi_{i,j}) - 1) \,. \tag{2.5}$$

Therefore, problem (2.3) 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

$$L(\pi, \lambda_1, \lambda_2) = \langle \pi(i,j), c(i,j) \rangle - \varepsilon \, \mathrm{Ent}(\pi) - \langle \lambda_1(i), \sum_j \pi_{i,j} - \rho_1(i) \rangle - \langle \lambda_2(j), \sum_i \pi_{i,j} - \rho_2(i) \rangle \,, \tag{2.6}$$

and we obtain taking variations

$$c(i,j) + \varepsilon \log(\pi_{i,j}) - \lambda_1(i) - \lambda_2(j) = 0 \,. \tag{2.7}$$

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

$$\pi_{ij} = e^{\lambda_1(i) + \lambda_2(j) - c(i,j)} = D_1 e^{-c(i,j)} D_2 \,, \tag{2.8}$$

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,

$$\textbf{Sinkhorn algorithm:} \quad \begin{cases} D_1^k = \mathbf{1}_n ./ (A D_2^{k-1}) \\ D_2^k = \mathbf{1}_m ./ (A^T D_1^k) \end{cases}, \tag{2.9}$$

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

$$\begin{cases} D_1^k = \rho_1 ./ (e^{-c/\varepsilon} D_2^{k-1}) \\ D_2^k = \rho_2 ./ ([e^{-c/\varepsilon}]^T D_1^k) \end{cases}. \tag{2.10}$$

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 1 (Discrete Entropic OT)**.**

$$\mathrm{OT}_\varepsilon(\rho_1, \rho_2) \stackrel{\text{def.}}{=} \min_{\pi \in \Pi(\rho_1, \rho_2)} \langle \pi(i,j), c(i,j) \rangle + \varepsilon \, \mathrm{KL}(\pi \,|\, \rho_1 \otimes \rho_2), \tag{2.11}$$

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

$$\mathrm{KL}(\rho \,|\, \mu) \stackrel{\text{def.}}{=} \sum_i \rho(i) \left(\log(\rho(i)/\mu(i)) - 1 \right). \tag{2.12}$$

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 1. *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* (2.11) *and* (2.3) *are the same. The formulation* (2.3) *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 1 (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.

2.1. **Convergence of Sinkhorn algorithm in the continuous setting.** As recalled in Fenchel-Rockafellar theorem 35, 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

$$y_{n+1} = \arg\max_y f(x_n, y) \tag{2.13}$$

$$x_{n+1} = \arg\max_x f(x, y_{n+1})\,. \tag{2.14}$$

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

Proposition 2 (Dual Problem)**.** *The dual problem reads* $\sup_{u,v} D(u,v)$ *where* $u, v \in C^0(X)$ *and*

$$D(u,v) = \langle u(x), \alpha(x)\rangle + \langle v(y), \beta(y)\rangle - \varepsilon\langle \alpha\otimes\beta, e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}} - 1\rangle\,. \tag{2.15}$$

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 3 (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*

$$u_{k+1}(x) = -\varepsilon\log\left(\int_X e^{\frac{v_k(y)-c(x,y)}{\varepsilon}}\,\mathrm{d}\beta(y)\right) (=: S_\beta(v_k)) \tag{2.16}$$

$$v_{k+1}(y) = -\varepsilon\log\left(\int_X e^{\frac{u_{k+1}(x)-c(x,y)}{\varepsilon}}\,\mathrm{d}\alpha(x)\right) (=: S_\alpha(u_{k+1}))\,. \tag{2.17}$$

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

$$1 - e^{u(x)/\varepsilon}\int_X e^{\frac{v(y)-c(x,y)}{\varepsilon}}\,\mathrm{d}\beta(y) = 0 \text{ for } x\,\alpha \text{ a.e.} \tag{2.18}$$

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 2 (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 4. *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 2.

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,

$$S_\beta(\tilde{v}) = \tilde{u} \tag{2.19}$$

$$S_\alpha(\tilde{u}) = \tilde{v}\,. \tag{2.20}$$

Formulas (2.19) (together with (2.18)) 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 5. *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,*

$$\|S_\alpha(u_1) - S_\alpha(u_2)\|_{\circ,\infty} \leq \kappa \|u_1 - u_2\|_{\circ,\infty} \tag{2.21}$$

where $\kappa < 1$ and where we define the norm in oscillation of f,

$$\|f\|_{\circ,\infty} \stackrel{\text{def.}}{=} \frac{1}{2}(\sup f - \inf f) = \inf_{a \in \mathbb{R}} \|f(x) - a\|_{\infty,\alpha}\,. \tag{2.22}$$

where the sup, inf and sup norm are taken w.r.t. α. Sometimes, we use $\operatorname{osc}(f) = (\sup f - \inf f)$.[2]

Proof. The first part of the proposition is obvious and used in the proof of Proposition 3. More precisely, the 1-Lipschitz property can be actually obtained by using

$$|S_\alpha(u_1)(x) - S_\alpha(u_2)(x)| = \left| \int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} S_\alpha(u_2 + t(u_1 - u_2))\,\mathrm{d}t \right| \tag{2.23}$$

$$\leq \int_0^1 \left| \int_X (u_1 - u_2) \frac{e^{\frac{t(u_1-u_2)}{\varepsilon}}}{\int_X e^{\frac{t(u_1-u_2)}{\varepsilon}} e^{\frac{u_2 - c(x,\cdot)}{\varepsilon}} \,\mathrm{d}\alpha} e^{\frac{u_2 - c(x,\cdot)}{\varepsilon}} \,\mathrm{d}\alpha \right| \mathrm{d}t \tag{2.24}$$

$$\leq \|u_1 - u_2\|_\infty\,. \tag{2.25}$$

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

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

point (an obvious case for balanced optimal transport). Using the same arguments as above, one has, for $u_1 \neq u_2$

$$\|S_\alpha(u_1) - S_\alpha(u_2)\|_{\circ,\infty} \leq \|S_\alpha(u_1) - S_\alpha(u_2)\|_\infty < \|u_1 - u_2\|_{\circ,\infty} \tag{2.26}$$

since the case of equality implies that $u_1 = u_2$. Refining the above inequality (2.25), one has

$$|S_\alpha(u_1)(x) - S_\alpha(u_2)(x)| \leq \kappa \|u_1 - u_2\|_{\circ,\infty}\,, \tag{2.27}$$

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

$$\kappa = \sup_{f\in\mathcal{S}\setminus\{0\}} \sup_{\tilde{v}\in\mathcal{V}} \frac{1}{\|f\|_\infty} \int_X f(x)\,\mathrm{d}\tilde{v}(x)\,, \tag{2.28}$$

where $\mathcal{V} \stackrel{\text{def.}}{=} \{\tilde{v} = \frac{1}{Z}e^V\,\mathrm{d}\alpha \ :\ V \in \frac{1}{\varepsilon}\mathcal{S}\}$ and Z is the normalizing constant to make $\tilde{v}$ 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 6 (Linear convergence of Sinkhorn)**.** *The sequence (u_k, v_k) linearly converges to (u_*, v_*) for the sup norm up to translation $\|\cdot\|_{\circ,\infty}$.*

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,

$$\|S_\beta \circ S_\alpha(u_1) - S_\beta \circ S_\alpha(u_2)\|_{\circ,\infty} \leq \kappa(\alpha)\kappa(\beta)\|u_1 - u_2\|_{\circ,\infty}\,, \tag{2.29}$$

therefore, the convergence is linear. □

Remark 3. *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 5 and we now give a quantitative estimate by a direct computational argument.

Proposition 7. *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 $S_\alpha(f)$:

$$\frac{1}{2}\,|S_\alpha(u_1)(y) - S_\alpha(u_2)(y) - S_\alpha(u_1)(x) - S_\alpha(u_2)(x)| \leq \frac{1}{2}\left|\int_0^1 \langle u_1 - u_2, \nu_{t,y} - \nu_{t,x}\rangle\,\mathrm{d}t\right|\,, \tag{2.30}$$

where $\nu_{t,z} \stackrel{\text{def.}}{=} \frac{1}{Z}e^{\frac{t(u_1-u_2)+u_2-c(z,\cdot)}{\varepsilon}}\,\mathrm{d}\alpha$ (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 8. Thus , we get

$$\|S_\alpha(u_1) - S_\alpha(u_2)\|_{\circ,\infty} \leq \kappa\|u_1 - u_2\|_{\circ,\infty}\,, \tag{2.31}$$

where κ is the constant estimated in Lemma 8 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 8. *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*

$$\|\nu_u - \nu_v\|_{L^1} \leq 2(1 - e^{-\|u-v\|_{\circ,\infty}}) = 2(1 - e^{-\frac{1}{2}\,\mathrm{osc}(u-v)})\,. \tag{2.32}$$

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

$$\psi_g'(t) + \psi_{v-u}(t)\psi_g(t) = \psi_{(v-u)g}(t)\,, \tag{2.33}$$

and therefore

$$e^{\int_0^t \psi_{v-u}(s)\,\mathrm{d}s}\psi_g(t) - \psi_g(0) = \int_0^t \psi_{(v-u)g}(s) e^{\int_0^s \psi_{v-u}(r)\,\mathrm{d}r}\,\mathrm{d}s\,. \tag{2.34}$$

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

$$|e^{\int_0^t \psi_{u-v}(s)\,\mathrm{d}s}(\psi_g(t) - \psi_g(0))| \leq |e^{\int_0^t \psi_{u-v}(s)\,\mathrm{d}s}\psi_g(t) - \psi_g(0)| \leq \|g\|_\infty \left(e^{\int_0^t \psi_{u-v}(s)\,\mathrm{d}s} - 1\right)\,. \tag{2.35}$$

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

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

and therefore, by optimizing[3] on the parameter t_0, we obtain

$$|\psi_g(1) - \psi_g(0)| \leq |\psi_g(1)| + |\psi_g(0)| \leq 2\|g\|_\infty (1 - e^{-\frac{1}{2}\int_0^1 \psi_{u-v}(s)\,\mathrm{d}s})\,. \tag{2.36}$$

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

$$\|\nu_u - \nu_v\|_{L^1} \leq 2(1 - e^{-\|u-v\|_{\circ,\infty}})\,. \tag{2.37}$$

□

Remark 4. *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.*

2.2. **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 2 (Hilbert metric)**.** Let $\mathbb{R}^n_{++}$ be the cone of positive coordinates vector. The Hilbert metric on this cone is

$$\mu(x,y) \stackrel{\text{def.}}{=} \max_{i,j} \log\left(\frac{x_i y_j}{x_j y_i}\right)\,. \tag{2.38}$$

A few remarks are in order: the quantity μ is nonnegative since one can take $i = j$ in Formula (2.38) 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

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

is simple to obtain and ensures that the Hilbert metric indeed is a metric on $[\mathcal{S}_n]_{++} \overset{\text{def.}}{=} \mathcal{S}_n \cap \mathbb{R}^n_{++}$, which is one possible parametrization of this quotient space. An important fact concerning the Hilbert metric is the following:

Theorem 9. *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 10. *Let $A \in \mathbb{R}^{m\times n}_{++}$ be a matrix with positive coefficients, then*

$$\mu(Ax, Ay) \leq \kappa(A)\mu(x, y) \forall x, y \in \mathbb{R}^n_{++} \tag{2.39}$$

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

$$\Delta(A) = \max_{i,j} \mu(Ae_i, Ae_j) = \max_{ijkl} \log\left(\frac{A_{ik}A_{jl}}{A_{il}A_{jk}}\right). \tag{2.40}$$

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:

Theorem 11. *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 12. *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 (2.38). □

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

Theorem 13. *The discrete Sinkhorn algorithm* (2.9) *linearly converges to its unique solution.*

Proof. Consider the sequences D_1^k and D_2^k generated by the Sinkhorn algorithm (2.9). One has

$$\mu(D_2^k, D_2^{k+1}) = \mu(\mathbf{1}_m./(A^T D_1^k), \mathbf{1}_m./(A^T D_1^{k+1})) = \mu(A^T D_1^k, A^T D_1^{k+1}) \leq \kappa(A^T)\mu(D_1^k, D_1^{k+1}). \tag{2.41}$$

Therefore, iterating this argument leads to

$$\mu(D_2^k, D_2^{k+1}) \leq \kappa(A)^2 \mu(D_2^k, D_2^{k-1}) \tag{2.42}$$

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 $\frac{2}{\varepsilon}LD >> 1$,

$$\Delta(A) \leq \frac{2}{\varepsilon}LD \text{ and } \kappa(A) \simeq (1 - e^{-\frac{1}{\varepsilon}LD})^2 \simeq 1 - 2e^{-\frac{1}{\varepsilon}LD}. \tag{2.43}$$

It can be compared with the constant we get in Proposition 7, $\kappa = 1 - e^{-\frac{1}{\varepsilon}L\,\mathrm{diam}(\alpha)}$. 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 8. 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.

2.3. **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 (2.9), 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 (2.9) is not available any longer and as a consequence, one cannot use optimized and parallelized implementations of matrix multiplication.

3. The Riemannian(-like) structure of W_2 and the Benamou-Brenier formula

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 discussion, we refer to [Santambrogio, 2015].

3.1. The primal problem of W_2 and its relaxation to the path space.

3.1.1. *Monge and Kantorovich.* Let us start with the Monge problem.

Definition 3 (Monge Problem)**.** Fix $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. Find a (measurable) map $T : X \mapsto Y$ s.t.

$$\inf_T \int_X c(x, T(x))\, \mathrm{d}\mu(x) \text{ s.t. } T_*\mu = \nu\,. \tag{3.1}$$

There are two issues with this formulation. The first one is that the optimization set may be empty since it is not possible to find a Monge pas that sends one Dirac to two Diracs of mass $1/2$. The second one is that the constraint is not convex making the problem difficult to solve. The famous relaxation proposed by Kantorovich addresses these issues. Informally, it allows for (infinitely) many maps which are associated to a given amount of mass at each location.

Definition 4 (Kantorovich Problem)**.** Fix $\mu \in \mathcal{P}(X)$ and $\nu \in \mathcal{P}(Y)$. A coupling plan for μ, ν is:

$$\gamma \in \mathcal{P}(X \times Y) \text{ s.t. } [p_1]_*\gamma = \mu \text{ and } [p_2]_*\gamma = \nu\,. \tag{3.2}$$

The set of coupling plans is denoted by $\mathcal{C}(\mu, \nu)$. The Kantorovich problem is the following linear programming (LP) problem

$$\mathrm{MK}_c(\mu, \nu) = \min_{\gamma \in \mathcal{C}(\mu,\nu)} \langle \gamma(x,y), c(x,y) \rangle\,. \tag{3.3}$$

From the mathematical point of view, existence of a minimizer is easy to establish under mild assumptions since the set of coupling plans $\mathcal{C}(\mu,\nu)$ is compact for weak-* convergence of measures (one has to show tightness and use Prokhorov theorem). If $c(x,y)$ is continuous[4], then existence of a minimizing coupling plan is granted.

The main question is why this convex problem is the correct relaxation of the Monge problem. It is possible to prove equality of these two quantities when μ has density with respect to the Lebesgue measure. Note that it is always true in a discrete setting in which empirical measures with uniform weights and equal number of points are used. Indeed, the Kantorovich problem is a linear optimization problem for which there exists a solution at an extreme point of the optimization set: The extreme points of this set are permutation matrices which defines a map between the points. If the cost is continuous and the sets X, Y are compact, it is then easy to approximate any measure with empirical measures with n distinct points, for which Monge maps are optimal.

Note that this LP has a particular form that can be exploited for numerical efficiency. When dealing with (uniform) empirical measures with the same weights, algorithms such as the Hungarian or auction with complexity in $O(n^3)$ can be used. When weights are not uniform, network flow algorithms are used with a complexity in $O(n^3\log(n))$.

3.1.2. *Kantorovich on the path space.* Let us assume that $X = Y = M$ is a manifold and that the cost is given by a Lagrangian

$$c(x,y) = \inf\left\{\int_0^1 L(x,\dot{x},\dots,x^{(n)})\,\mathrm{d}t\,;\ x\in C^n([0,1],M) \text{ and } (x(0),x(1)) = (x,y)\right\}, \tag{3.4}$$

where $\dot{x},\dots,x^{(n)}$ denote the time derivatives of the path x. The most important example being the case of the Riemannian squared distance on (M,g) being a Riemannian manifold: $d^2(x,y)$ on $M\times M$ where d its associated Riemannian distance. For instance, in the case M is a Riemannian manifold with a metric g, one can consider the induced distance squared

$$c(x,y) = \inf\left\{\int_0^1 g_x(\dot{x},\dot{x})\,\mathrm{d}t\,;\ x\in C^1([0,1],M) \text{ and } (x(0),x(1)) = (x,y)\right\}. \tag{3.5}$$

The Kantorovich optimization set is a probability on pairs of points, however, this cost arises from an infimum problem on the set of paths. For instance, on a complete Riemannian manifold where there is no cut locus, implying that for any two points there exists a unique geodesic between two points, to a given pair of points, corresponds a unique minimizing path for the kinetic energy. One can try to relax a bit further the Kantorovich problem by looking at a probability measure on $C([0,1],M)$ and a cost $c(\omega)$ for $\omega\in C([0,1],M)$. Then, one can form the following optimization problem:

$$\inf_{\pi}\langle\pi(\omega),c(\omega)\rangle \tag{3.6}$$

under some marginal constraints. In constrast with the Kantorovich problem, since the evalution map $e_{t_1,\dots,t_n}:\omega\mapsto(\omega(t_1,\dots,\omega(t_m))$ is continuous, one can impose multiple marginal constraints at different times $t_1,\dots,t_n$. For the squared distance cost, the cost is $c(\omega) = \int_0^1 g_x(\dot\omega,\dot\omega)\,\mathrm{d}t$ and $+\infty$ if it does not exist. Then, one retrieves the Kantorovich two marginal constraints via $e_{t=0}$ and $e_{t=1}$. Assuming the existence of an optimal plan π on the path space, we get an optimal plan for the Kantorovich problem via pushforward $[e_{t=0,t=1}]_\star\pi$. It also gives a way to interpolate in time the two densities μ and ν by $[e_t]_\star(\pi)$.

This time interpolation for the case of the quadratic cost in euclidean spaces (or Riemannian manifold complete and empty cut locus), is simple: particules follow straight lines. Once an optimal potential φ has been found between a density (wrt Lebesgue) μ and a measure ν, the time dependent map is $\varphi_t = \mathrm{Id} + t(\nabla\varphi - \mathrm{Id})$ so that $[e_t]_\star(\pi) = [\varphi_t]_\star\mu$. However, for a different cost on the path space such as an acceleration cost (e.g. $\|\ddot\omega\|^2$), interpolations are more complicated.

[4]Actually, lower semicontinuous and bounded below is sufficient.

3.2. **Monge on the path space and Benamou-Brenier.** 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 as in the paragraph above. It can be seen as a reformulation of the problem on the path space as a control problem on the space of time-dependent measures ρ_t. Let us first introduce this formulation and then discuss it.

Definition 5 (Benamou-Brenier - non-convex formulation)**.** Consider a Riemannian manifold (M, g),

$$\inf_{\rho,v} \int_0^1 \int_M g(v(t,x), v(t,x))\, \mathrm{d}\rho(x)\, \mathrm{d}t\,, \tag{3.7}$$

under the continuity equation constraint $\partial_t \rho(t,x) + \mathrm{div}(\rho(t,x)v(t,x)) = 0$ and time boundary constraints $\rho(0) = \rho_0, \rho(1) = \rho_1$.

In order to introduce it, we remark that the Monge formulation on the path space has not been yet introduced. If instead of finding a plan, one look for a path of maps φ_t that solves the Monge problem, we have

$$\int_X \frac{1}{2}\, d^2(x, \varphi_1(x))\, \mathrm{d}\mu(x) \leq \int_0^1 \int_X \frac{1}{2}\, g(\varphi(t,x))(\partial_t\varphi(t,x), \partial_t\varphi(t,x)) d\rho_0(x) dt\,, \tag{3.8}$$

for every path φ_t of maps such that $\varphi_0 = \mathrm{Id}$ and φ_1 that pushforward ρ_0 onto ρ_1. In the case of the Euclidean cost, the equality is achieved if and only if $\varphi_t = \mathrm{Id} + t(\nabla\varphi_1 - \mathrm{Id})$. Therefore, optimization on paths of maps reduces to optimization of the Monge problem. Using the time-dependent change of variable $y = \varphi(t,x)$, we get

$$\int_0^1 \int_X \frac{1}{2} g(\varphi(t,x))(\partial_t\varphi(t,x), \partial_t\varphi(t,x)) d\rho_0(x) dt = \frac{1}{2}\int_0^1 \int_X g(y)(v(t,y), v(t,y)) d\rho_t(x) dt \tag{3.9}$$

by definition of the image measure. This equality is simply the rewriting of the Lagrangian expressed in Lagrangian coordinates (parametrize by a moving particle) into Eulerian (parametrize by a fixed point in space) coordinates.

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. It is one of the key contribution by Benamou and Brenier to propose a convex reformulation amenable to non-smooth convex optimization:

Definition 6 (Benamou-Brenier - convex formulation)**.** Consider two given measures $\rho_0, \rho_1 \in \mathcal{P}(M)$ and the optimization set $m \in \mathcal{M}([0,1] \times M, TM)$ i.e. measure taking values in the tangent space of M and $\rho(t,x) \in \mathcal{M}([0,1] \times M)$.

$$\frac{1}{2}\inf_{\rho,m} \int_0^1 \int_M P_{\|x\|^2}(m(t,x), \rho(t,x))\, \mathrm{d}\rho(t,x)\, \mathrm{d}t\,, \tag{3.10}$$

under the linear constraint $\partial_t\rho(t,x) + \mathrm{div}(m) = 0$ and same time boundary constraints on ρ. The linear constraint is satisfied in the following weak sense, for every $f(t,x) \in C^1([0,1] \times M)$,

$$-\langle \partial_t f, \rho\rangle - \langle \nabla f, m\rangle = \langle f(t=1), \rho_1\rangle - \langle f(t=0), \rho_0\rangle\,. \tag{3.11}$$

On a closed Riemannian manifold, there is no boundary conditions but on a bounded convex set in $\mathbb{R}^d$, this weak constraint also encodes the homogeneous Neumann boundary condition (zero flux). Proving existence of minimizers can be done through the use of Fenchel-Rockafellar by first defining the (pre-)dual problem. The result can be guessed since the functional is positively one-homogeneous, it Legendre transform is a convex indicator which appears as a constraint: Namely, $\varphi \in C$ where $C = \{\varphi; (\partial_t\varphi(t,x), \nabla\varphi(t,x)) \in D, \forall (t,x) \in [0,1] \times M\}$ where $D = \{a + \frac{\|b\|^2}{2} \leq 0\} \subset \mathbb{R} \times \mathbb{R}^d$ and the dual objective function is linear in ρ_0 and ρ_1

$$\sup_{\varphi \in C} \langle \varphi(t=1), \rho_1\rangle - \langle \varphi(t=0), \rho_0\rangle\,. \tag{3.12}$$

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. An explicit proof on Riemannian manifolds including source terms in the continuity equation can be found in []. We give hereafter a strategy of proof.

Sketch of proof on $\mathbb{R}^d$. Use a mollifier and convolution on $\rho(t,x)$ to define a smooth $\rho^\varepsilon(t,x)$ for a smoothing parameter ε. It implies that $\rho^\varepsilon(t,x)$ satisfies the same continuity equation for the momentum m^ε which is the convolution between the chosen mollifier and m. In particular, one can define a smooth velocity[5] v^ε. Then, integrate the flow of this smooth velocity field, write it in Lagrangian coordinates to obtain the Monge formulation. By one homogeneity and convexity Its value is less than the objective functional evaluated at ρ, m. However, the boundary conditions are lost. However, convolution with a mollifier, for instance gaussian, can be explicitly bounded in Wasserstein. It finishes to prove that it is the correct relaxation of the Monge problem. □

This formulation was introduced by Benamou and Brenier for numerical purposes. Indeed, one can apply non-smooth convex optimization algorithms to solve the optimization problem in Definition 6.

Note that the BB formulation only involves kinetic energy; it is also possible to introduce a potential energy on the space of densities, such as the Fisher information

$$V(\rho) = \int_M |\nabla(\log\rho)|^2 \,\mathrm{d}\rho(x) = \int_M \frac{|\nabla\rho(x)|^2}{\rho(x)} dx\,,$$

where the second inequality is correct when ρ is sufficiently smooth. However, this energy is also a convex functional of ρ since it is also the pointwise integration of a perspective function with arguments $(\nabla\rho, \rho)$. As a consequence, it can be extended from smooth densities to more general measures. It is used in one of the formulation of the Schrödinger problem.

3.3. **The Wasserstein Riemannian(-like) metric.** 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 $\mathcal{P}(M)$ with an additional variable which is the velocity field. We first rewrite the cost in the optimal transport functional on the space of vector fields: that is, if $\rho_1 = (\exp\varepsilon v)_*(\rho_0)$ 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 v defined on M. For instance, on the Euclidean space, assuming that the coupling is $\pi_\varepsilon = (\mathrm{id}, \mathrm{id} + \varepsilon v)_*\rho_0$, we get

$$\langle \pi_\varepsilon, d(x,y)^2\rangle \simeq \varepsilon^2 \int_M \|v(x)\|^2 \,\mathrm{d}\rho_0(x)\,. \tag{3.13}$$

Our goal is to identify the Riemannian(-like) metric of optimal transport. We consider a tangent vector to a probability density ρ (say smooth and positive) that we denote by $\delta\rho$. Then, to find the norm of $\delta\rho$, it is sufficient to find the vector field that will infinitesimally pushforward ρ onto $\rho + \varepsilon\delta\rho$. The action of a vector field on a density is $-\operatorname{div}(\rho v) = \delta\rho$. Note that there are many vector fields that can reproduce this infinitesimal change in ρ. The selection of v is done by minimization of the kinetic functional

$$\inf_v \frac{1}{2}\int \|v\|^2 d\rho(t,x) \text{ under the constraint } -\operatorname{div}(\rho v) = \delta\rho\,. \tag{3.14}$$

Using Lagrange multipliers, the vector field v is necessarily the gradient of Lagrange multiplier. Therefore, one can replace the space of vector fields by the space of function p, however, in this case the constraint reads

$$\Delta_\rho(p) := -\operatorname{div}(\rho\nabla p) = \delta\rho\,. \tag{3.15}$$

On a closed (connected) Riemannian manifold and if ρ has sufficient (mild) regularity, this is an elliptic equation whose solution is unique up to a constant. Now, one can write the Riemannian(-like) metric tensor of the Wasserstein space, at least for sufficiently smooth densities ρ:

[5]Notice that this velocity is NOT the convolution by the mollifier of the corresponding vector field v.

Proposition 14. *The Riemannian(-like) metric tensor at a density ρ is*

$$\int_M \|\nabla \Delta_\rho^{-1} \delta\rho\|^2 d\rho(x)\,. \tag{3.16}$$

which is also equal to (when dealing with L^2 densities)

$$\int_M \Delta_\rho^{-1}(\delta\rho)\delta\rho dx\,. \tag{3.17}$$

The order of the Wasserstein metric is -1 since one first integrate twice and then differentiate once in the above formula. It is thus similar to an H^{-1} metric that depends on the current density. It is tempting to compare the Wasserstein metric and a form of H^{-1} metric (the formula above in which ρ is taken to be constant as the reference volume measure, e.g. Lebesgue in Euclidean space). It seems an easy computation if there exists a lower bound and an upper bound on the density. However, the key point is to control that quantity along geodesics. Such types of results have been proven in the literature and the key point of this section is that the Wasserstein metric is locally like an H^{-1} type metric, which might be simpler to compute if one is just interested in the metric tensor. It is the case for gradient flows which are discussed next.

Remark that when ε is sufficiently small, $|x|^2/2 + \varepsilon \nabla F(x)$ is convex when ∇F is smooth enough.

3.4. **Gradient flows.** Gradient flows with respect to the Wasserstein metric is now a well-known and well-studied subject. This literature was triggered by Otto's work on the porous medium equation [Otto, 2001] and the famous JKO (Jordan-Kinderlehrer-Otto) scheme in [Jordan et al., 1998]. We briefly present it now from an informal point of view since it is connected with convex optimization. Of course, it is the most important example of application where one actually does not need the full optimal transport structure (i.e. computing geodesics), but only the Riemannian(-like) structure. First recall how to compute a gradient in the case of a function on the Euclidean space with respect to a metric defined by a symmetric positive matrix A. It has the following variational formulation (check using first-order optimality condition):

$$\nabla f(x) = \arg\min_w \frac{1}{2}\langle w, Aw\rangle - df_x(w)\,,$$

where the dot product is with respect to an ambient L^2 metric. It gives $\nabla f(x) = A^{-1}(df_x)$. We simply need to apply this formula on the space of densities. We use this formulation in a similar way for the Wasserstein space. Consider a functional on the space of densities denoted by $F(\rho)$, which is Fréchet differentiable. We denote by $\frac{\delta F}{\delta \rho}$ its derivative. By the formula above (3.17) the operator $A = \Delta_\rho$ and the gradient flow is

$$\partial_t \rho = -\Delta_\rho\left(\frac{\delta F}{\delta \rho}\right) = \operatorname{div}\left(\rho \nabla \frac{\delta F}{\delta \rho}(\rho)\right)\,. \tag{3.18}$$

Let us perform the same computation again but at the level of vector fields instead of using directly the metric tensor (actually this is a very similar derivation but we do it again for pedagogical purpose). In mathematical terms, using the same approach than in Formula (3.14), we get

$$\arg\min_v \frac{1}{2}\int_M \|v(t,x)\|^2 \,\mathrm{d}\rho(x) - \left\langle \frac{\delta F}{\delta \rho}(\rho), -\operatorname{div}(\rho v)\right\rangle\,, \tag{3.19}$$

where we informally denoted by $\frac{\delta F}{\delta \rho}$ the Fréchet derivative of F. Remark that it is not a direct application of the formula for the gradient recalled above since the minimization is done at the level of vector fields. We get now, $v = \nabla \frac{\delta F}{\delta \rho}(\rho)$ and thus, again

$$\partial_t \rho = \operatorname{div}\left(\rho \nabla \frac{\delta F}{\delta \rho}(\rho)\right)\,. \tag{3.20}$$

3.4.1. ***The Fokker-Planck equation with a strongly convex potential V.*** The Fokker-Planck equation is the PDE that arises from the law of the process associated with the Langevin dynamic:

$$dX(t) = -\nabla V(X(t)) + \sqrt{2} dB(t)\,. \tag{3.21}$$

The law of $X(t)$ denoted by $\rho(t)$ evolves accordingly to

$$\partial_t \rho = \mathrm{div}(\rho \nabla V) + \Delta \rho\,. \tag{3.22}$$

It is known that the density $\rho(t)$ converges to $\rho_\infty = \frac{1}{\int e^{-V(x)} dx} e^{-V}$. Note that ρ_∞ is the unique steady state of this PDE. Indeed, $0 = \mathrm{div}(\rho \nabla V) + \Delta \rho = \mathrm{div}(\rho(\nabla V + \nabla \log \rho)$ which implies $\rho = e^{-V + cste}$ and the constant is chosen so that ρ is a probability measure.

Importantly, this process is used in Bayesian approaches in order to simulate under the law of interest ρ_∞. The key point in simulating (3.21) is that one does not need to have access to the renormalizing constant $\int e^{-V(x)} dx$. Indeed, the flow equation is invariant to the addition of a constant to the potential.

It appears that the Wasserstein geometry is suitable to study the convergence of $\rho(t)$ to ρ_∞ which is known to hold. This convergence will be proven next using this point of view. Let us apply the gradient flow computation to the entropy+potential functional $F(\rho) = \int_X \rho(x)(\log(\rho(x)) - 1)\,\mathrm{d}x + \int_X V(x)\rho(x)\,\mathrm{d}x$ for which $\frac{\delta F}{\delta \rho}(\rho) = \log(\rho) + V(x)$ and so

$$\partial_t \rho = \Delta \rho + \mathrm{div}(\rho \nabla V)\,, \tag{3.23}$$

which is again the Fokker-Planck equation. Now, the functional F enjoys geodesic convexity: the linear term is convex in the Wasserstein space if and only if V is convex. Let us prove this result: geodesics between two Diracs are geodesics in M. So geodesic convexity in the Wasserstein space implies plain (geodesic) convexity of V on M, so the convexity of V is necessary. It is also sufficient since any geodesic curve in the Wasserstein space is supported by geodesics (for instance via the Lagrangian representation on the path space) on M. Consequently, if V is convex, then it is also the case of $\rho \mapsto \int_M V(x) d\rho(x)$. We thus have the following fact:

Proposition 15. *The term $\int_M V(x) d\rho(x)$ is geodesically convex in Wasserstein, if and only if V is (geodesically[6]) convex. The entropy is also convex for $M = \mathbb{R}^d$.*

We now prove the second point: it can be seen by writing with $f(s) = s\log(s)$ and using the change of variable formula

$$\int_{\mathbb{R}^d} f(\rho(t,x)) dx = \int_{\mathbb{R}^d} f\Big(\frac{\rho_0(x)}{\det(\nabla \varphi(t,x))}\Big) \det(\nabla \varphi(t,x)) dx = \int_{\mathbb{R}^d} \log\Big(\frac{\rho_0(x)}{\det(\nabla \varphi(t,x))}\Big) dx\,. \tag{3.24}$$

Using the structure of $\varphi(t)$, $\varphi(t) = \mathrm{Id} + t(\varphi(1) - \mathrm{Id})$ and the fact that $\log(\det)$ is a concave function on the space of symmetric positive matrices, we get the result. An important result by McCann [McCann, 1997] gives sufficient conditions on a function f so that $\int_{\mathbb{R}^d} f(\rho) dx$ is convex in Wasserstein. Among the important examples is $f(x) = x^p$ for $p > 1$. The general conditions are that f is convex, $f(0) = 0$, superlinear and the following function is convex decreasing:

$$s \mapsto s^d f(s^{-d})\,. \tag{3.25}$$

Convexity of the entropy in Wasserstein on more general Riemannian manifold has been the topic of intense study which was pioneered by Lott, Villani [Lott and Villani, 2006] and Sturm [Sturm, 2006]. They first proved that the entropy is λ-convex if the Ricci curvature of the space is lower bounded. Then, λ-convexity of the entropy along W_2 geodesic can be taken as a definition for Ricci curvature on spaces that are more general than Riemannian manifolds.

As a consequence, the Kullback-Leibler divergence between ρ and ρ_∞ satisfies the PL inequality. Such an inequality is also called a log-Sobolev inequality and it reads

$$\mathrm{KL}(\rho, \rho_\infty) \leq \frac{1}{2\lambda} \int_{\mathbb{R}^d} \|\nabla \log(\rho/\rho_\infty)\|^2 d\rho(x) = \frac{1}{2\lambda} \|\nabla_\rho \mathrm{KL}(\rho, \rho_\infty)\|^2_{W^2}\,. \tag{3.26}$$

[6] On a complete Riemannian manifold.

The consequence of Proposition 15 and Corollary 31 is that, for a constant C

$$\mathrm{KL}(\rho, \rho_\infty) \leq Ce^{-2\lambda t}. \tag{3.27}$$

Exponential convergence with respect to KL has been obtained in a simple and elegant manner. We have seen that interpreting some particular PDE's from the point of view of gradient flows allows for simple analysis. However, this allows to extend the corresponding PDE to general measures instead of densities.

3.4.2. ***A word on the time-discrete scheme: the JKO scheme.*** This line of research originates from De Giorgi's minimizing movement scheme [de Giorgi, 1993]. One can now define implicit gradient scheme similar to Definition 19 by replacing the Hilbert norm with the Wasserstein distance, with τ a timestep parameter,

$$\rho_{k+1} = \arg\min_\rho \frac{1}{2\tau} W_2^2(\rho_k, \rho) + F(\rho). \tag{3.28}$$

The convergence of this time discrete scheme in the case of entropy has been proven by Jordan, Kinderlehrer and Otto [Jordan et al., 1998] and a rather complete study of this scheme is given in Ambrosio, Gigli, Savaré's book on Gradient Flows [Ambrosio et al.,]. A lighter introduction is given in Santambrogio's "{Euclidean, metric, and Wasserstein} gradient flows" [Santambrogio, 2016]. We refer the reader to these two last references for a detailed discussion of gradient flows.

Remark 5. *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 divergence on the space of densities for which the underlying metric tensor is the same than the Wasserstein distance would be suitable.*

Remark 6. *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. It also gives a practical numerical scheme to implement this PDE in discrete time, with interesting properties such as preserving the positivity constraint.*

3.5. **Unbalanced optimal transport (UOT).** Following the Benamou and Brenier formulation, there has been lots of models proposed in the literature deriving from it. We focus hereafter on the so-called Wasserstein-Fisher-Rao model [Chizat et al., 2015], which is also called Hellinger-Kantorovich [Liero et al., 2015].

3.5.1. *Dynamic formulation of unbalanced optimal transport.* We choose to present the extension of optimal transport to unbalanced optimal transport, that is optimal transport with creation/deletion 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

$$\partial_t \rho + \mathrm{div}(\rho v) = \alpha\rho \tag{3.29}$$

where we introduced a source term parametrized by the growth rate α which depends both on time and space. Then, we have to postulate[7] a Lagrangian on this growth rate and a natural action for this is the Fisher-Rao functional

$$\frac{1}{2}\int_M \alpha^2 \,\mathrm{d}\rho. \tag{3.30}$$

[7]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.

With this Lagrangian, the extension of the Benamou-Brenier formulation is as follows, minimize, under Equation (3.29), the action

$$\inf_{\rho,v,\alpha} \int_0^1 \int_M \frac{1}{2}(\|v(t,x)\|^2 + \frac{1}{4}\alpha(t,x)^2)\,\mathrm{d}\rho(t,x)\,\mathrm{d}t\,, \tag{3.31}$$

where we emphasized the dependence of the control variable on time and space. This slight modification of Benamou-Brenier leads to the following dual problem: Maximize $\int_M \varphi_1(y)d\rho_1(y) - \int_M \varphi_0(y)d\rho_0(y)$ under the constraint that $\varphi \in C^1([0,1],M)$ such that

$$\partial_t\varphi + \frac{1}{2}(\|\nabla\varphi\|^2 + \varphi^2) \leq 0\,. \tag{3.32}$$

Note that, as it is the case in standard OT, this equation does not depend on the current density $\rho(t)$. It implies that given one $\varphi(t)$ that realizes the equality (instead of the inequality), one can use it to solve the generalized continuity equation and thus obtain a path of minimizing energy.

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. Using the language of convex analysis, the new metric tensor is obtained as the infimal convolution of the Wasserstein metric tensor and the Fisher-Rao metric tensor, whence its name Wasserstein-Fisher-Rao. It has been named Hellinger-Kantorovich in [Liero et al., 2015] since at the level of distances, it interpolates between Kantorovich and Hellinger distances. Something that is not clear from this dynamic formulation is the existence of a Kantorovich formulation of the problem. There are different ways to discover the existence of an associated Kantorovich problem:

(1) In standard OT, the Hamilton-Jacobi equation has known solutions: Hopf-Lax solutions which are infinimal convolution of the initial φ_0 and the distance squared. One would need an equivalent of such solution, which is however is hard to find in the literature. We do not discuss it further.
(2) A more pedestrian way 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 the variables position and mass, it reads

$$d^2((x,m),(y,n)) = m + n - 2\sqrt{mn}\cos(\min(d(x,y),\frac{\pi}{2}))\,, \tag{3.33}$$

which is one-homogeneous in the mass variables (m,n).

One key property of the standard Wasserstein distance is that it is positively 1-homogeneous, with respect to mass scaling. Recall the following property, whose proof is straightforward,

Lemma 16. *If $f : E \to \mathbb{R}$ is a convex function on a vector space E which is positively homogeneous, then f is sub-additive, namely*

$$f(x+y) \leq f(x) + f(y)\,. \tag{3.34}$$

A direct consequence of this property (W_p is the p-Wasserstein distance) is that

$$W_p^p\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \frac{1}{n}\sum_{i=1}^n \delta_{y_i}\right) \leq \frac{1}{n}\sum_{i=1}^n W_p^p(\delta_{x_i},\delta_{y_i})\,. \tag{3.35}$$

This inequality uses a particular decomposition of the measures and it applies the homogeneity and convexity property. In fact, optimizing on all possible decomposition of the measures, one retrieves, in the case of probability measures, the standard definition of the Wasserstein distance as a linear programming problem.

Note that the dynamic formulation of UOT leads to a convex problem on the space of positive Radon measures, which is also positively homogeneous. In particular, it is natural to transfer such a formulation to UOT. It is the purpose of the next section to convince the reader that it can be done.

3.5.2. *Kantorovich formulation of WFR.* A priori, formulating an optimization problem in which one can decompose the marginals freely in collections of particles in position and mass and optimizing over it, should give an upper bound on the dynamic formulation problem because the continuity equation in momentum variables is linear. There is a slight technical problem in formulating such an optimization problem since one has to deal with the possibility of having 0 mass at some points, i.e. to deal with the apex of the cone. A way to overcome this issue is to consider a Dirac at this point with 0 mass which simply represents an absence of mass: $\delta_{(x,0)}$ for x in the base space. More generally, any Dirac mass on M can be represented via its lift on the cone by $(x,m) \mapsto \delta_{(x,m)}$. However, such a lift should be coherent with the structure of convex cone of positive measures. For instance the representation of a marginal using $p_1\delta_{(x,m_1)} + p_2\delta_{(x,m_2)}$ should be equivalent to the representation $\delta_{(x,m_1p_1+m_2p_2)}$. It simply means that one can represent a marginal μ on the M as the integration over a mass variable of a positive measure on the cone $\tilde{\mu}$ that satisfies:

$$\int_{\mathbb{R}_{\geq 0}} m d\tilde{\mu}(x,m) = \mu(x)\,. \tag{3.36}$$

The minimization problem reads

$$\min_{\tilde{\mu},\tilde{\nu}} \langle d^2((x,m),(y,n)), \tilde{\pi}((x,m),(y,n))\rangle \tag{3.37}$$

under the two moment constraints corresponding to the marginal constraints,

$$\begin{cases} \int_{\mathbb{R}_{\geq 0}} m d\tilde{\mu}(x,m) = \mu(x) \\ \int_{\mathbb{R}_{\geq 0}} n d\tilde{\nu}(x,n) = \nu(x)\,. \end{cases} \tag{3.38}$$

Using convexity and 1-homogeneity of the dynamic formulation, it can be proven that the two problems coincide. The main argument consists in smoothing the quantities to obtain a competitor that one can control. Now, let us write the dual problem associated with this moment constrained optimal transport on the cone. The main difference with a general cost, is the one homogeneity of the cost. Let us denote by p_0, p_1 the Lagrange multipliers associated with the marginal constraints and write

$$\sup_{p_0,p_1\in C(X)} \langle mp_0(x) + np_1(y), \pi_1 \otimes \pi_2\rangle = \langle p_0(x) + p_1(y), \mu\otimes\nu\rangle \tag{3.39}$$

under the constraint $mp_0(x) + np_1(y) \leq d((x,m),(y,n))^2$. Observe that the constraint is a 2-homogeneous polynomial function in terms of $r = \sqrt{m}$ and $s = \sqrt{n}$, so that one can use the discriminant of this polynomial function to obtain

$$\begin{aligned} &(1-p_0(x))r^2 + (1-p_1(y))s^2 - 2rs\cos(\min(d(x,y),\pi)) \geq 0 \iff \\ &\qquad \cos(\min(d(x,y),\pi))^2 \leq (1-p_0(x))(1-p_1(y)) \text{ and } (1-p_i(x)) \geq 0 \text{ for} i = 1,2\,. \end{aligned} \tag{3.40}$$

Pay attention to the fact that this polynomial function is only observed on $\mathbb{R}^2_{>0}$ and not on the whole $\mathbb{R}^2$. However, the dominating coefficients are positive by constraint and the derivative (of the polynomial function in terms of r/s) at 0 is nonpositive. As a consequence, nonnegativity of this polynomial function on $\mathbb{R}^2_{>0}$ implies its nonnegativity on $\mathbb{R}^2$.
The key point is that the constraint can be transformed into a constraint similar to standard optimal transport by simply taking the logarithm. With $h_0 = -\log(1-p_0(x))$ and $h_1 = -\log(1-p_1(x))$, the objective functional becomes

$$\sup_{h_0,h_1} \langle 1 - e^{-h_0}, \mu\rangle + \langle 1 - e^{-h_1}, \nu\rangle\,, \tag{3.41}$$

under the constraint

$$h_0(x) + h_1(y) \leq -\log(\cos(\min(d(x,y),\pi))^2)\,. \tag{3.42}$$

Fortunately, taking this logarithm change of variable, the convexity of the problem is not destroyed, which is a remarkable fact. The objective functional resembles to the Legendre transform of the

entropy $x\log(x) - x + 1$ so that it could be understood as the Legendre transform of an integral functional such as Kullback-Leibler.

Exercise: Write the dual formulation using the Fenchel-Rockafellar theorem (see Appendix, Theorem 35) and compare with the formula proposed below.

The dual formulation of the above problem with variables h_0, h_1 results in a reduced problem on the space of positive Radon measures: Given μ, ν two positive 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),

$$WFR(\mu,\nu)^2 = \inf_{\pi\in\mathcal{M}_+(M\times M)} \mathrm{KL}(\pi_1,\mu) + \mathrm{KL}(\pi_2,\nu) + \langle \pi, -\log(\cos^2(\min(d(x,y),\frac{\pi}{2})))\rangle\,, \tag{3.43}$$

where the optimization is performed on π which is a positive Radon measure on the product space $M \times M$. For the rigorous proof of this theorem, we refer the reader to [Chizat et al., 2015] or [Liero et al., 2015]. The surprising fact in the Kantorovich formulation above is the cost which appears in the scalar product and it is explained by the derivation above. 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 (3.43), therefore it shows the importance of the WFR formulation (this fact is proven in [Liero et al., 2015]). A general approach is developed in [Liero et al., 2015] that provides a recipe to go from a static formulation in a similar form than (3.43) with different Csizar divergences to the conic formulation.

From the point of view of numerics however, the conic formulation is a priori less attractive than the reduced formulation (3.43) since it has an additional dimension in the mass variable. Yet both formulations are amenable to entropic regularization and lead to different regularization. The formulation that is most commonly used in practice is the reduced formulation (3.43) with an associated Sinkhorn algorithm for which linear convergence can be proven also for more general divergence terms.

3.5.3. *An informal discussion of Kantorovich formulation in position and mass.* One could directly start by postulating a cost in position and mass, say (x, m) and solve an optimal transport problem in this variable. As previously explained, one issue are the marginal constraints in this "conic" formulation which are not defined. The lift from the space of positive measures to positive measures on the cone can be explicitly defined by

$$m\delta_x \mapsto \delta_{x,m}\,. \tag{3.44}$$

and for a density

$$\rho(x) \mapsto \delta_{\rho(x)}(r)\otimes dx\,. \tag{3.45}$$

However, in general, the measure on the right-hand side has infinite total mass when working with a reference measure on the base which has infinite volume. One can try to correct this point by introducing another reference measure, say η that is a probability density and use

$$m\delta_x \mapsto \delta_{x,m}\,. \tag{3.46}$$

and for a density

$$\rho(x) \mapsto \delta_{\rho(x)/\mu(x)}(r)\otimes d\eta(x)\,. \tag{3.47}$$

However, one must then define an optimal transport problem on the cone that does not depend on the choice of this reference density. Let us give it a try and write the Kantorovich formulation for such a transport problem with reference density μ. We have

$$\sup_{f,g} \int_{M\times\mathbb{R}_{\geq 0}} f(x,m)\delta_{\mu(x)/\eta(x)}(m)\otimes d\eta(x) + \int_{M\times\mathbb{R}_{\geq 0}} g(y,n)\delta_{\nu(x)/\eta(x)}(n)\otimes d\eta(x)\,. \tag{3.48}$$

and the constraint $f(x,m)+g(y,n) \leq c((x,m),(y,n))$. A sufficient condition for the resulting problem to be independent of the choice of the reference density (in the case of measures with densities on the base space), is to impose test functions f and g to be one-homogeneous with respect to the second variable. As seen above, imposing such a constraint on test functions is implied by imposing moment constraints with respect to the mass variable. However, it implies that a pre-optimization can be done on the free choice of reference measure η which results in replacing the cost c by $\tilde{c}_{x,y}(m,n) := \inf_{a>0} c((x,am),(y,an))/a$, thus turning the cost into a one-homogeneous function in (m,n).

Now, the inequality constraint reads as, denoting $f(x,m) = m\tilde{f}(x)$ and similarly for g,

$$m\tilde{f}(x) + n\tilde{g}(y) \leq \tilde{c}((x,m),(y,n))\,. \tag{3.49}$$

In terms of Legendre transform in (m,n), this inequality can be written $\tilde{c}^*_{x,y}(\tilde{f}(x),\tilde{g}(y)) \leq 0$. One-homogeneity of the cost $\tilde{c}$ implies that $\tilde{c}^*_{x,y}$ is the convex indicator function of a convex set denoted by $C(x,y)$. The resulting problem becomes

$$\sup_{\tilde{f},\tilde{g}} \int_M \tilde{f}(x)d\mu(x) + \int_M \tilde{g}(y)d\nu(y)\,, \tag{3.50}$$

under the constraint

$$(\tilde{f}(x),\tilde{g}(y)) \in C(x,y)\,. \tag{3.51}$$

Remark that the resulting problem is completely expressed in terms of quantities which do not make use of the cone construction. From this discussion, it is not difficult to prove that if the cost on the cone is one-homogeneous in (m,n) and a power of a distance on the cone, then the corresponding problem defines a power of a distance on the space of positive Radon measures. The main argument of the proof of the triangle inequality is to use the homogeneity property and the standard gluing lemma.
What is left open in our discussion, is the link with the primal problem (3.43) that was formulated using Csizar divergences on the marginals.

3.5.4. *Metric structure and examples of gradient flows.* Based on a similar computation in the OT case, it is not difficult to write the metric tensor on the space of nonnegative smooth function which are integrable, at least very formally.

Exercise 1: Write the Riemannian(-like) metric tensor corresponding to the Wasserstein-Fisher-Rao metric.
Exercise 2: Write the Wasserstein-Fisher-Rao gradient flow of the entropy: $\int(\rho(x)\log(\rho(x)) - 1)dx$.

What is important is to find functionals that are geodesically convex for this new metric, akin to the entropy+potential functional in the classical OT case. This question have been addressed in [Laschos and Mielke, 2022].

3.5.5. *Dynamic formulation of entropic regularization.* Let us make a small detour to classical entropic regularization in optimal transport. Importantly, the entropic regularization has also a dynamic formulation on the space of densities. One has the equality

$$\mathrm{OT}_\varepsilon(\rho_0,\rho_1) + \frac{d\varepsilon}{2}\log(2\pi\varepsilon) = -\frac{\varepsilon}{2}(\mathrm{KL}(\rho_0,\mathrm{Leb}) + \mathrm{KL}(\rho_1,\mathrm{Leb})) + \inf_{\rho,v} \int_0^1 \int_{\mathbb{R}^d} \left(\frac{1}{2}|v|^2 + \frac{\varepsilon^2}{8}|\nabla\log(\rho)|^2\right) \mathrm{d}\rho\,\mathrm{d}t\,, \tag{3.52}$$

under the continuity equation constraint $\partial_t\rho + \mathrm{div}(\rho v) = 0$. The term in $-\int |\nabla \log(\rho)|^2 d\rho$ 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

$$\begin{cases} \partial_t\rho + \mathrm{div}(\rho\nabla p) = 0 \\ \partial_t p + \frac{1}{2}|\nabla p|^2 = \frac{\delta}{\delta\rho}\left(\frac{\varepsilon^2}{4}\int_M |\nabla\log(\rho)|^2)\,\mathrm{d}\rho\right). \end{cases} \tag{3.53}$$

Interestingly, this system can be transformed by introducing the following change of variables $z = p - \frac{\varepsilon}{2}\log(\rho)$[8] to get

$$\begin{cases} \partial_t\rho + \mathrm{div}(\rho\nabla z) = \frac{\varepsilon}{2}\Delta\rho \\ \partial_t z + \frac{1}{2}|\nabla z|^2 = -\frac{\varepsilon}{2}\Delta z\,. \end{cases} \tag{3.54}$$

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 (3.60) 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.

Using the following change of variables (Hopf-Cole formula)

$$\begin{cases} \eta(t,x) = \sqrt{\rho(t,x)}e^{z(t,x)/\varepsilon} \\ \eta^\star(t,x) = \sqrt{\rho(t,x)}e^{-z(t,x)/\varepsilon}\,, \end{cases} \tag{3.55}$$

this system is transformed into

$$\begin{cases} \partial_t\eta(t,x) = \frac{\varepsilon}{2}\Delta\eta \\ \partial_t\eta^\star(t,x) = -\frac{\varepsilon}{2}\Delta\eta^\star\,. \end{cases} \tag{3.56}$$

Each of these equations have explicit solutions in Euclidean space which is given by convolution with a Gaussian kernel. This fact explains why it is possible to obtain a static formulation of the corresponding dynamic formulation of Schrödinger bridge. Note that the corresponding density is given by $\eta(t,x)\eta^\star(t,x)$.

3.5.6. *Entropic regularization and UOT.* We now use the geometric perspective of Léger [Léger, 2017] to introduce the corresponding formulation of a Schrödinger bridge for Wasserstein-Fisher-Rao which was recently proposed in [Buze and Duong, 2023] with a rather different argument. The Schrödinger problem can be abstracted (although without sharing all the nice properties) to the following form. Consider a Riemannian manifold (M,g), a function $F : M \to \mathbb{R}$ and the following Lagrangian, defined on curves $C^1_{pcw}([0,1],M)$

$$\int_0^1 g(x(t))(\dot{x}(t),\dot{x}(t)) + \frac{1}{4}g(x)(\nabla F(x(t)),\nabla F(x(t)))dt\,. \tag{3.57}$$

Remark that it is related the Schrödinger problem for $M = \mathcal{P}_1(\mathbb{R}^d)$, the metric g being the Wasserstein metric and the function F being the entropy. The curve $\rho(t) \in \mathcal{P}(\mathbb{R}^d)$ is subject to the continuity equation $\partial_t\rho(t) + \mathrm{div}(v\rho(t)) = 0$, the corresponding length (not optimized over v) is $\int_{\mathbb{R}^d}|v(x)|^2 d\rho(t,x)$ and the squared norm of gradient of the entropy is the Fisher information $\int_{\mathbb{R}^d}|\nabla\log(\rho(t,x))|^2 d\rho(t,x)$.

The point in considering this formulation is to make clear the relation between two formulations: rewrite

$$\dot{x} = u - \frac{1}{2}\nabla F(x)$$

for u a tangent vector at x. Then, optimize on u the Lagrangian

$$\int g(x(t))(u,u)dt \tag{3.58}$$

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

subject to the previous constraint. In the Wasserstein case with entropy as function F, the constraint reads

$$\partial_t \rho + \mathrm{div}(\rho u) = \frac{1}{2}\Delta\rho$$

and the Lagrangian reads

$$\int_{\mathbb{R}^d} |v(x)|^2 d\rho(t,x)\,.$$

Now, expand the squares,

$$g(x)(\dot{x} + \frac{1}{2}\nabla F(x), \dot{x} + \frac{1}{2}\nabla F(x)) = g(x(t))(\dot{x}(t), \dot{x}(t)) + \frac{1}{4}g(x)(\nabla F(x(t)), \nabla F(x)) + g(x)(\nabla F(x), \dot{x})$$

Remark that the last term can be explicitly integrated

$$\int_0^1 g(x)(\nabla F(x), \dot{x})dt = \int_0^1 DF(x)(\dot{x})dt = F(x(1)) - F(x(0))\,.$$

Therefore, the two problems of minimization are equivalent up to boundary terms. In the Wasserstein case, it is the entropy of $\rho(t=0)$ and $\rho(t=1)$. The same computation holds for the Wasserstein-Fisher-Rao metric: what is different here is the metric and therefore gradients. For instance, considering the entropy $\mathrm{Ent}(\rho) = \int \rho\log(\rho) - \int \rho$, its gradient norm squared with respect to Wasserstein leads to $\int |\nabla \log(\rho)|^2 d\rho(x)$ whereas with respect to Wasserstein-Fisher-Rao, it leads to $\int |\nabla \log(\rho)|^2 + \frac{1}{\delta^2}\log(\rho)^2 d\rho(x)$. The system of equation for the entropy and the Wasserstein-Fisher-Rao metric leads to

$$\partial_t \rho(t) + \mathrm{div}(\rho u) - \frac{1}{2}\Delta\rho = \alpha\rho - \frac{1}{2\delta^2}\rho\log(\rho)\,. \tag{3.59}$$

and minimization of the kinetic energy associated with WFR:

$$\int_0^1 \int_{\mathbb{R}^d} (|u(t,x)|^2 + \delta^2\alpha(t,x)^2) d\rho(t,x)dt\,.$$

In an other formulation,

$$\inf_{\rho,v,\alpha} \int_0^1 \int_{\mathbb{R}^d} \left(\frac{1}{2}|v|^2 + \frac{\delta^2}{2}|\alpha(t,x)|^2 + \frac{\varepsilon^2}{8}(|\nabla\log(\rho)|^2 + \frac{1}{\delta^2}|\log(\rho(t,x))|^2\right) \mathrm{d}\rho\,\mathrm{d}t\,. \tag{3.60}$$

under the constraint that

$$\partial_t \rho + \mathrm{div}(\rho u) = \alpha\rho\,. \tag{3.61}$$

What is left open in our discussion is the existence of a static model associated with this formulation. Not known in the current literature is a probabilistic model associated with it.

4. 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 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,

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

(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 17. *Define*

$$S_\varepsilon(\mu,\nu) = OT_\varepsilon(\mu,\nu) - \frac{1}{2}(OT_\varepsilon(\mu,\mu) + OT_\varepsilon(\nu,\nu))\,. \tag{4.1}$$

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. Using the formulation (3.60), it is possible to justify, at least formally, why this formula is expected to lead to better approximation of optimal transport.

4.1. **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

$$W_2(\rho_0,\rho_1)^2 = \inf_{\rho,m} K(\rho,m) + \iota_C(\rho,m)\,. \tag{4.2}$$

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 4.1. 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[9]. 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:

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

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

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,

$$\min_{(\rho,m,\tilde\rho,\tilde m)} K(\rho,m)+\iota_C(\tilde\rho,\tilde m)+\iota_{\text{interp}}((\rho,m),(\tilde\rho,\tilde m)) \tag{4.3}$$

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 (4.3). 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 $G_2=\iota_{\text{interp}}$. 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

$$\min_x \frac{1}{2}\|x-x_0\|^2+\frac{1}{2}\|Ax-y_0\|^2\,. \tag{4.4}$$

Optimality implies

$$x-x_0+A^*(Ax-y_0)=0\,, \tag{4.5}$$

and thus

$$x=(\mathrm{Id}+A^*A)^{-1}(A^*y_0+x_0)\,. \tag{4.6}$$

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

$$A(\tilde\rho,\tilde m)=\begin{pmatrix}\mathrm{div}(\tilde\rho,\tilde m)\\ s_{BC}(\tilde\rho,\tilde m)\end{pmatrix}=\begin{pmatrix}0\\ b_0\end{pmatrix}\,, \tag{4.7}$$

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

$$\mathrm{prox}_{\iota_C}(z)=\arg\min_x \frac{1}{2}|x-z|^2 \text{ s.t. } Ax=\begin{pmatrix}0\\ b_0\end{pmatrix}\,. \tag{4.8}$$

Using Lagrange multipliers, the optimality condition leads to

$$x=z+A^*p \tag{4.9}$$

$$Ax=Az+AA^*p=\begin{pmatrix}0\\ b_0\end{pmatrix}\,. \tag{4.10}$$

which implies

$$AA^*p = \begin{pmatrix} 0 \\ b_0 \end{pmatrix} - Az\,. \tag{4.11}$$

Remark 7. *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,

$$x = z + A^*(AA^*)^{-1}(v_0 - Az)\,. \tag{4.12}$$

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

$$\frac{1}{2}|x - p_{\overline{BC}}(x_0)|^2 + \langle p, Ax - Ab_0\rangle \tag{4.13}$$

and get

$$\mathrm{div}\, p_{\overline{BC}} p^*_{\overline{BC}}\, \mathrm{div}^*\, p = Ab_0 - Ax_0\,, \tag{4.14}$$

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

$$\arg\min_{\rho,m} \frac{1}{2\tau}|m_0 - m|^2 + \frac{1}{2\tau}|\rho - \rho_0|^2 + \frac{1}{2}\frac{|m|^2}{\rho}\,. \tag{4.15}$$

Variations in m and ρ lead to

$$\frac{1}{2\tau}(m - m_0) + \frac{m}{\rho} = 0 \tag{4.16}$$

$$\frac{1}{2\tau}(\rho - \rho_0) + \frac{1}{2}\frac{|m|^2}{\rho^2} = 0\,. \tag{4.17}$$

These two equations imply the two following relations

$$m = \frac{m_0}{(1 + \frac{2\tau}{\rho})} \tag{4.18}$$

and

$$(\rho + 2\tau)^2(\rho - \rho_0) - \tau\rho^2|m_0|^2 = 0\,. \tag{4.19}$$

By uniqueness of the proximal map, the argmin is the unique (if it exists) positive root of Equation (4.19). 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 8. *In fact, $\frac{|m|^2}{\rho}$ being one-homogeneous, the Legendre-Fenchel conjugate is the convex indicator*

$$C_0 := \{(\alpha, \beta)\,|\,\alpha + \frac{1}{2}|\beta|^2 \leq 0\}\,. \tag{4.20}$$

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

REFERENCES

[Ambrosio et al.,] Ambrosio, L., Gigli, N., and Savaré, G. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics ETH Zürich. Birkhäuser, 2. ed edition. OCLC: 254181287.

[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*.

[Buze and Duong, 2023] Buze, M. and Duong, M. H. (2023). Entropic regularisation of unbalanced optimal transportation problems.

[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.

[de Giorgi, 1993] de Giorgi, E. (1993). New problems on minimizing movements.

[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.

[Jordan et al., 1998] Jordan, R., Kinderlehrer, D., and Otto, F. (1998). The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17.

[Laschos and Mielke, 2022] Laschos, V. and Mielke, A. (2022). Evolutionary variational inequalities on the hellinger-kantorovich and spherical hellinger-kantorovich spaces.

[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.

[Lott and Villani, 2006] Lott, J. and Villani, C. (2006). Ricci curvature for metric-measure spaces via optimal transport.

[Léger, 2017] Léger, F. (2017). A geometric perspective on regularized optimal transport.

[McCann, 1997] McCann, R. J. (1997). A convexity principle for interacting gases. *Advances in Mathematics*, 128:153–179.

[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.

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

[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.

[Santambrogio, 2016] Santambrogio, F. (2016). Euclidean, Metric, and Wasserstein gradient flows: an overview.

[Sturm, 2006] Sturm, K.-T. (2006). On the geometry of metric measure spaces. *Acta Mathematica*, 196(1):65 – 131.

APPENDIX A. 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.

A.1. **Usual definitions.**

Definition 7. 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 8. A function $f : E \mapsto [-\infty, \infty]$ is convex if its epigraph defined as

$$\text{epi}(f) \stackrel{\text{def.}}{=} \{(x, y) : y \geq f(x)\} \subset E \times \mathbb{R} \tag{A.1}$$

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

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

$$f(x) \leq \lim_{n\to\infty} f(x_n)\,. \tag{A.2}$$

Example 18. *Let $C \subset E$ be a set. We denote by $\iota_C : E \mapsto \mathbb{R}$ the indicator function of C defined as*

$$\iota_C(x) = \begin{cases} 0 \text{ if } x \in C\,, \\ +\infty \text{ otherwise.} \end{cases} \tag{A.3}$$

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

$$\min_{x\in C} f(x) = \min_{x\in C} f(x) + \iota_C(x)\,. \tag{A.4}$$

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

Proposition 19 (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 20 (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 21. *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 10 (Convex conjugate)**.** Let $f : E \mapsto \mathbb{R}$ be a function. The convex conjugate $f^* : E^* \mapsto \mathbb{R}$ is defined as

$$f^*(p) = \sup_{x\in E} \langle p, x\rangle - f(x)\,. \tag{A.5}$$

Proposition 22. *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 three very important examples of convex functions that will be used heavily in optimal transport formulation. The first example has already be implicitly given:

Definition 11. Let $f : E \mapsto \mathbb{R}$ be a convex function and proper. The Bregman divergence of the function is

$$B_f(y; x) := f(y) - \langle \nabla f(x), y - x\rangle - f(x)\,. \tag{A.6}$$

This function is convex wrt to y and nonnegative. If in addition f is strictly convex, $B_f(y; x) = 0$ iff $y = x$.

In general, Bregman divergences are not convex with respect to the second argument. There is however one notable exception that we consider as an example.

Example 23. *Consider $f(x) = x\log(x)$, then $B_f(y;x) = y\log(y) - (\log(x)+1)(y-x) - x\log(x) = y\log(y/x) - y + x$. Although it is easily convex in both x, y, we will give another general proof using the perspective function defined below.*

The second example is the Legendre transform of a convex set.

Proposition 24 (Positive one-homogeneous convex functions and convex indicator functions are in convex duality)**.** *Let C be a convex set in E. Then, ι_C^* is a convex, positively one-homogeneous function. Reciprocally, given a convex, positively one-homogeneous function, its Legendre transform is the convex indicator of a convex set.*

Proof. Indeed, for a positive real λ, $\iota_C^*(\lambda p) = \sup_x \langle \lambda p, x\rangle - \iota_C(x)$. For a given x, whether $\iota_C(x) = +\infty$ and in this case $\langle \lambda p, x\rangle - \iota_C(x) = \lambda(\langle \lambda p, x\rangle - \iota_C(x))$. In the other case, $\iota_C(x) = 0$, and here again $\langle \lambda p, x\rangle - \iota_C(x) = \lambda(\langle \lambda p, x\rangle) = \lambda(\langle \lambda p, x\rangle - \iota_C(x))$. Taking the supremum gives the first result.

For the second point, we have consider remark that taking the supremum on x equals taking the supremum in λx for positive λ and x. Remark that $\langle p, \lambda x\rangle - f(\lambda x) = \lambda(\langle p, x\rangle - f(x))$, which after optimization on λ gives two cases: whether there exists x such that $\langle p, x\rangle - f(x)) > 0$ and in this case,$f^*(p) = +\infty$, whether $\langle p, x\rangle - f(x)) \leq 0$ for every x. The set C of all p satisfying this inequality for all x is convex and in this case, it implies that the optimization on x is obtained for $x = 0$, in which case $\langle p, 0\rangle - f(0)) = 0$. We have proven that $f^* = \iota_C$. □

The third example is the construction of a perspective function. For this new construction, we need the notion of the recession function.

Definition 12 (Recession function)**.** Let $f : E \mapsto (-\infty, +\infty]$ be a convex proper and lsc function. Its recession function is defined by, for a given y such that $f(y) < +\infty$,

$$\mathrm{rec}_f(x) \coloneqq \lim_{t\to\infty} \frac{f(y+tv)}{t}\,. \tag{A.7}$$

It does not depend on the chosen y.

Definition 13 (Perspective function)**.** Let $f : E \mapsto (-\infty, +\infty]$ be a convex proper and lsc function. The perspective function $P_f : E \times \mathbb{R} \mapsto (-\infty, +\infty]$ is defined as

$$P_f(x,s) = \begin{cases} sf(x/s) \text{ if } s > 0\,, \\ \mathrm{rec}_f(x) \text{ if } s = 0\,, \\ +\infty \text{ otherwise.} \end{cases}\,. \tag{A.8}$$

This function is convex, lsc and proper. It is also one homogeneous wrt (x, s).

The name perspective function maybe comes the fact that its graph wrt x gives different scaled version of the graph of f.

Example 25. *Some important examples are the following:*

1. *$f(x) = \|x\|^2$, its recession function is $+\infty$ and its perspective function is $(x,s) \to \frac{\|x\|^2}{s}$.*
2. *$e(x) = x\log(x) - x + 1$. In addition to being convex, $e(0) = 1$ and $e(x) \geq 0$ since $e(x) = B_{x\log(x)}(x, 1)$. Its perspective function is $P_e(x,s) = x\log(x/s) - x + s = B_{x\log(x)}(x;s)$.*
3. *Let h be a positively one-homogeneous convex function, then $P_f(x,s) = f(x)$.*

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

Definition 14 (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

$$\partial f(x) \stackrel{\text{def.}}{=} \{p \in E^* : f(y) \geq f(x) + \langle p, y - x\rangle \text{ for all } y \in E\}\,. \tag{A.9}$$

Remark 9. *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 26. *The definition of subgradient implies, exchanging the order of x, y in the inequality* (A.9) *and adding the two inequalities*

$$\langle \partial f(x) - \partial f(y), x - y \rangle \geq 0\,, \tag{A.10}$$

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

Proposition 27 (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 10. *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 15 (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 28 (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*

$$\langle \nabla f^*(x) - \nabla f^*(y), x - y \rangle \geq \lambda \|\nabla f^*(x) - \nabla f^*(y)\|^2\,, \tag{A.11}$$

∇f is a co-coercive monotone operator.

A.2. **Elementary convex optimization.**

Definition 16 (PL condition)**.** Let (M, g) be a Riemannian manifold (possibly of infinite dimension) and $F : M \mapsto \mathbb{R}$ which is C^1 with at least one global minimizer $x_\star \in M$. Then, F satisfies a Polyak-Lojasiewicz inequality if

$$F(x) - F(x_\star) \leq C\|\nabla F(x)\|^2 \tag{A.12}$$

for a positive constant C.

Remark that if the PL condition implies that every local minimizer is a global one since $\nabla F(x) = 0$ implies $F(x) - F(x_\star) = 0$. Note that the PL condition can be weakened to a local PL condition by making the constant C dependant on a given ball $C(r)$. Such functions also satisfies that local minimizers are global.

Definition 17 (Gradient flow)**.** Let $f : M \mapsto \mathbb{R}$ be a C^1 function. The gradient flow associated with f is

$$\dot{x} = -\nabla f(x)\,, \tag{A.13}$$

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

Interestingly, this notion can be defined using weaker formulations, in particular in metric space settings, see Gradient flows, Ambrosio, Gigli and Savaré. Moreover, the PL condition only requires to measure the norm of the gradient, which is well defined in the optimal transport context.

Proposition 29. *If F satisfies the PL condition, then the (continuous) gradient flow satisfies*

$$F(x(t)) - F(x_\star) \leq e^{-t/C}(F(x_0) - F(x_\star))\,, \tag{A.14}$$

where $x(t)$ satisfies $\dot{x} = -F(x(t))$.

Proof.

$$\frac{d}{dt}(F(x(t)) - F(x_\star)) = -\|\nabla F(x(t))\|^2 \leq -\frac{1}{C}(F(x(t)) - F(x_\star)), \tag{A.15}$$

which implies the result by Grönwall's lemma. □

Note that this condition does not need any convexity, neither on the function nor on the optimization set since it applies to Riemannian manifolds. Obviously, the task is to show that the PL condition is satisfied. A sufficient condition to ensure it is to be in a convex optimization setting.

Proposition 30. *If $F : H \mapsto \mathbb{R}$ is λ-strongly convex, then F satisfies the PL condition for the constant $\frac{1}{2\lambda}$.*

Proof. Write the strong convexity inequality at point x, so that

$$F(x_\star) - F(x) \geq \langle \nabla F(x), x_\star - x\rangle + \frac{\lambda}{2}\|x - x_\star\|^2. \tag{A.16}$$

Reversing the sign of the inequality and using the standard inequality $\langle x, y\rangle \leq \frac{1}{2\alpha}\|x\|^2 + \frac{\alpha}{2}\|y\|^2$ for $\alpha = \lambda$, we get

$$F(x) - F(x_\star) \leq \frac{1}{2\lambda}\|\nabla F(x)\|^2 + \frac{\lambda}{2}\|x - x_\star\|^2 - \frac{\lambda}{2}\|x - x_\star\|^2 = \frac{1}{2\lambda}\|\nabla F(x)\|^2. \tag{A.17}$$

□

Remark that the proof also holds on a Riemannian manifold, so we get:

Corollary 31. *A λ-strongly convex function on a Riemannian manifold (M, g) satisfies the PL condition.*

Definition 18 (Explicit gradient descent)**.** A time-discrete counterpart of the gradient flow is an explicit formulation (the gradient is computed at the current point) with constant step size gradient descent, for $\tau > 0$,

$$x_{k+1} = x_k - \tau\nabla f(x_k). \tag{A.18}$$

Proposition 32. *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

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + L/2\|y - x\|^2, \tag{A.19}$$

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

$$f(x_{k+1}) \leq f(x_k) + \tau\langle \nabla f(x_k), \nabla f(x_k)\rangle + L\tau^2/2\|\nabla f(x_k)\|^2 \tag{A.20}$$

$$\leq \tau(-1 + L\tau/2)\|\nabla f(x_k)\|^2. \tag{A.21}$$

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

$$x_{k+1} = x_k - \tau\nabla f(x_{k+1}). \tag{A.22}$$

This time-discrete scheme has a variational formulation,

$$x_{k+1} = \arg\min \frac{1}{2\tau}\|x - x_k\|^2 + f(x)\,, \tag{A.23}$$

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).

Remark that the variational formulation might still be convex even when f is not convex for τ sufficiently small since the convexity of $\frac{1}{2\tau}\|x - x_k\|^2$ can be sufficiently strong to compensate the lack of convexity of f. This argument does not apply when f has some (concave) singularity, e.g. $f(x) = -\|x\|$.

Proposition 33. *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 20. Let f be a convex function, proper and lsc. The proximal operator is defined as

$$\mathrm{prox}_{\tau f}(x) = \arg\min_y \frac{1}{2\tau}\|x - y\|^2 + f(y)\,. \tag{A.24}$$

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

$$\mathrm{prox}_{\tau f}(x) - x + \tau\partial f(x) \ni 0\,. \tag{A.25}$$

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

In particular, if it is reasonably cheap to compute the proximal operator of f, then the implicit gradient descent $x_{k+1} = \mathrm{prox}_{\tau f}(x_k)$ 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 34. *Let f be a convex, proper and lsc function. Then, it holds*

$$x = \mathrm{prox}_{\tau f}(x) + \tau\,\mathrm{prox}_{\frac{1}{\tau}f^*}(\frac{1}{\tau}x)\,, \tag{A.26}$$

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

$$\min_x f(x) + g(x)\,, \tag{A.27}$$

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

$$f(x) + g(x) \leq f(x) + g(x_*) + \langle \nabla g(x_*), x - x_*\rangle + \frac{L}{2}\|x - x_*\|^2\,, \tag{A.28}$$

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

$$x_{k+1} = \mathrm{prox}_{(1/L)f}(x_k - \frac{1}{L}\nabla g(x_k))\,, \tag{A.29}$$

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 (A.27). 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

$$\min_x f(Kx) + g(x)\,, \tag{A.30}$$

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 (A.30).

$$\min_x \max_p \langle p, Kx\rangle - f^*(p) + g(x) \geq \max_p \min_x \langle p, Kx\rangle - f^*(p) + g(x) \tag{A.31}$$

$$\geq \max_p -g^*(-K^*p) + f^*(p)\,, \tag{A.32}$$

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 35 (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*

$$\sup_{x\in E} -f(-x) - g(Lx) = \min_{p\in F^*} f^*(L^*p) + g^*(p)\,. \tag{A.33}$$

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.

A.3. **Primal-dual.** The problem of interest consists in the minimization of

$$\inf_x f(Kx) + g(x) \tag{A.34}$$

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

$$\inf_x \max_p \langle p, Kx\rangle - f^*(p) + g(x)\,. \tag{A.35}$$

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

$$0 \in K^*p + \partial g(x) \tag{A.36}$$

which can be alternatively rewritten as

$$x - \tau K^*p \in (\mathrm{Id} + \tau\partial g)(x)\,. \tag{A.37}$$

Writing a similar equation on p leads to

$$x \leftarrow (\mathrm{Id} + \tau_1\partial g)^{-1}(x - \tau_1 K^*p) \tag{A.38}$$

$$p \leftarrow (\mathrm{Id} + \tau_2\partial f^*)(p + \tau_2 Kx)\,, \tag{A.39}$$

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

$$x_{k+1} \leftarrow \mathrm{prox}_{\tau_1 g}(x_k - \tau_1 K^*p) \tag{A.40}$$

$$p_{k+1} \leftarrow \mathrm{prox}_{\tau_2 f^*}(p_k + \tau_2 K(2x_{k+1} - x_k))\,, \tag{A.41}$$

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.

A.4. **Augmented Lagrangian and ADMM.** Hereafter, the objective functions are of the type

$$\min_{Ax+By=b} f(x) + g(y)\,. \tag{A.42}$$

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,

$$\min_{x,y}\sup_{z} f(x) + g(y) + \langle z, b - Ax - By\rangle + \frac{\gamma}{2}\|b - Ax - By\|^2\,. \tag{A.43}$$

Then, the ADMM algorithm reads

$$x_{k+1} \leftarrow \arg\min_x f(x) - \langle z_k, Ax\rangle + \frac{\gamma}{2}\|b - Ax - By_k\|^2 \tag{A.44}$$

$$y_{k+1} \leftarrow \arg\min_y g(y) - \langle z_k, By\rangle + \frac{\gamma}{2}\|b - Ax_{k+1} - By\|^2 \tag{A.45}$$

$$z_{k+1} \leftarrow z_k + \gamma(b - Ax_{k+1} - By_{k+1})\,. \tag{A.46}$$

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

A.5. **Douglas-Rachford algorithm.** This algorithm is designed for the minimization of

$$\min_x g(x) + f(x) \tag{A.47}$$

one writes

$$x_{k+1} \leftarrow \mathrm{prox}_{\tau_1 g}(x_k - \tau_1 p_k) \tag{A.48}$$

$$p_{k+1} \leftarrow \mathrm{prox}_{\tau_2 f^*}(p_k + \tau_2(2x_{k+1} - x_k))\,, \tag{A.49}$$

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

$$x_{k+1} \leftarrow \mathrm{prox}_{\tau g}(v_k) \tag{A.50}$$

$$v_{k+1} \leftarrow v_k - x_{k+1} + \mathrm{prox}_{\tau f}(2x_{k+1} - v_k)\,. \tag{A.51}$$

Appendix B. Strong convexity of semi-dual Sinkhorn functional

In this section, we study convexity properties of the following functional:

$$S_\lambda(f) = \langle f, \mu\rangle - \lambda\langle \log\left(\int e^{\frac{f-c}{\lambda}}\,\mathrm{d}\mu\right), \nu\rangle s\,, \tag{B.1}$$

which is the dual objective function of the entropic regularized optimal transport optimized on the potential associated with the measure ν. The first derivative reads

$$S_\lambda(f)(\delta f) = \langle \delta f, \mu\rangle - \langle \delta f, \int \frac{e^{\frac{f-c}{\lambda}}}{\int e^{\frac{f-c}{\lambda}}\,\mathrm{d}\mu}\,\mathrm{d}\nu\rangle\,. \tag{B.2}$$

Therefore, the gradient of S_λ is

$$\nabla S_\lambda(f) = \mu - \int \frac{e^{\frac{f-c}{\lambda}}}{\int e^{\frac{f-c}{\lambda}}\,\mathrm{d}\mu}\,\mathrm{d}\nu\,. \tag{B.3}$$

It is convenient to introduce the measure $f[\nu] := \int \frac{e^{\frac{f-c}{\lambda}}}{\int e^{\frac{f-c}{\lambda}}\,\mathrm{d}\mu}\,\mathrm{d}\nu$ to rewrite the gradient as

$$\nabla S_\lambda(f) = \mu - f[\nu]\,. \tag{B.4}$$

The Hessian reads

$$\langle \delta f, \nabla^2 S_\lambda(f)\delta f\rangle = \frac{1}{\lambda}\langle (\delta f)^2, \int \frac{e^{\frac{f-c}{\lambda}}}{\int e^{\frac{f-c}{\lambda}}\,\mathrm{d}\mu}\nu\rangle - \frac{1}{\lambda}\langle \delta f, \int \frac{e^{\frac{f-c}{\lambda}}}{\int e^{\frac{f-c}{\lambda}}\,\mathrm{d}\mu}\nu\rangle^2 . \tag{B.5}$$

which can be rewritten as

$$\langle \delta f, \nabla^2 S_\lambda(f)\delta f\rangle = \frac{1}{\lambda}\,\mathrm{Var}_{f[\nu]}(\delta f) . \tag{B.6}$$

Since the potentials are defined up to an additive constant, we can deduce at least with no quantitative bounds that for all $f \in \mathrm{Lip}(X)$ such that $\mathrm{Lip}(f) \leq M$,

$$\mathrm{Var}_{f[\nu]}(\delta f) \geq cste\|\delta f\|^2_{\mathrm{osc}} . \tag{B.7}$$

since the quantity $\mathrm{Var}_{f[\nu]}(\delta f)$ achieves its lower bound on the compact subspace $C(X)/\mathbb{R}$ of Lipschitz functions. However, the constant needs to be estimated quantitatively.