Lecture 2: Entropic Optimal Transport

Luca Nenna

February 23, 2022

Introduction: Discret Optimal Transport

We now consider the optimal transport problems between probability measures on two finite sets X and Y with, for simplicity, both of cardinality N and we set

$$\mu = \sum_{x\in X} \mu_x \delta_x \qquad \nu = \sum_{y\in Y} \nu_y \delta_y.$$

Definition 0.1 (Discrete OT)**.** The discrete Optimal transport problem between two given measures μ and ν and a given cost function $c : X \times Y \to \mathbb{R}_+ \cup \{+\infty\}$ is the following minimization problem

$$\inf \left\{ \sum_{x\in X}\sum_{y\in Y} \gamma_{xy} c(x,y) \mid \gamma \in \Pi(\mu,\nu) \right\}, \tag{0.1}$$

where the set of admissible couplings is now defined as

$$\Pi(\mu,\nu) := \{\gamma \in X \times Y \mid \gamma_{xy} \geqslant 0,\ \sum_{y\in Y} \gamma_{xy} = \mu_x\ \forall x \in X,\ \sum_{x\in X} \gamma_{xy} = \nu_y\ \forall y \in Y\}.$$

Unfortunately, this linear programming problem has complexity $O(N^3)$ which actually means that it is infeasible for large N. A way to overcome this difficulty is by means of the **Entropic Regularization** which provides an approximation of Optimal Transport with lower computational complexity and easy implementation.

References: Entropic regularisation of Optimal Transport is a very active research field. We refer the interested reader to [1, 5, 9, 11, 6] and the citations therein. We also remark that these notes are inspired by the graduate classe on Numerical Optimal Transport given by F.-X. Vialard [13].

1 The Entropic Optimal Transport

1.1 The discrete case

We start from the primal formulation of the optimal transport problem, but instead of imposing the constraints $\gamma_{xy} \geqslant 0$, we add a term $\mathrm{Ent}(\gamma) = \sum_{x,y} e(\gamma_{xy})$, involving the (opposite of the) entropy

$$e(r) = \begin{cases} r(\log r - 1) & \text{if } r > 0 \\ 0 & \text{if } r = 0 \\ +\infty & \text{if } r < 0 \end{cases}$$

More precisely, given a parameter $\varepsilon > 0$ we consider

$$\boxed{P_\varepsilon = \inf\left\{\langle\gamma|c\rangle + \varepsilon\,\mathrm{Ent}(\gamma) \mid \gamma \in X \times Y,\ \sum_{y\in Y}\gamma_{xy} = \mu_x,\ \sum_{x\in X}\gamma_{xy} = \nu_y\right\},} \tag{1.2}$$

where $\langle\gamma|c\rangle = \sum_{x,y}\gamma_{xy}c(x,y)$ and $\mathrm{Ent}(\gamma) = \sum_{x,y} e(\gamma_{xy})$.

Theorem 1.1. *The problem P_ε has a unique solution $\gamma^\star$, which belongs to $\Pi(\mu,\nu)$. Moreover, if $\min(\min_{x\in X}\mu_x, \min_{y\in Y}\nu_y) > 0$ then*

$$\gamma_{x,y} > 0\ \forall (x,y) \in X \times Y.$$

Before introducing the duality, it is important to state the following convergence result in ε.

Theorem 1.2 (Convergence in ε). *The unique solution γ_ε to (1.2) converges to the optimal solution with minimal entropy within the set of all optimal solutions of the Optimal Transport problem, that is*

$$\gamma_\varepsilon \xrightarrow[\varepsilon\to 0]{} \operatorname{argmin}\left\{\mathrm{Ent}(\gamma) \mid \gamma \in \Pi(\mu,\nu),\ \langle\gamma|c\rangle = \mathcal{MK}_c(\mu,\nu)\right\}. \tag{1.3}$$

Proof. Consider a sequence $(\varepsilon_k)_k$ such that $\varepsilon_k \to 0$ and $\varepsilon_k > 0$ and denote γ_k the solution to (1.2) with $\varepsilon = \varepsilon_k$. Since $\Pi(\mu,\nu)$ is bounded and close we can extract a converging subsequence $\gamma_k \to \gamma^\star \in \Pi(\mu,\nu)$. Take now any optimal γ for the unregularized problem then by optimality of γ_k and γ one has

$$0 \leqslant \langle\gamma_k|c\rangle - \langle\gamma|c\rangle \leqslant \varepsilon_k(\mathrm{Ent}(\gamma) - \mathrm{Ent}(\gamma_k)). \tag{1.4}$$

Since $\mathrm{Ent}(\cdot)$ is continuous, by taking the limit $k \to +\infty$ in (1.4) we get $\langle\gamma^\star|c\rangle = \langle\gamma|c\rangle$. Furthermore, dividing by ε_k and taking the limit we obtain that $\mathrm{Ent}(\gamma) \geqslant \mathrm{Ent}(\gamma^\star)$ showing that $\gamma^\star$ is a solution to the minimization problem in (1.3). By strict convexity of Ent the optimization problem (1.3) has a unique solution and the whole sequence is converging to $\gamma^\star$. □

We want now to derive formally the dual problem. For this purpose we introduce the Lagrangian associated to (1.2)

$$\begin{aligned}\mathcal{L}(\gamma,\varphi,\psi) := \sum_{x,y}\gamma_{xy}c(x,y) + \varepsilon e(\gamma_{xy}) + \sum_{x\in X}\varphi(x)\left(\mu_x - \sum_{y\in Y}\gamma_{xy}\right)\\ + \sum_{y\in Y}\psi(y)\left(\nu_y - \sum_{y\in Y}\gamma_{xy}\right),\end{aligned} \tag{1.5}$$

where $\varphi : X \to \mathbb{R}$ and $\psi : Y \to \mathbb{R}$ are the Lagrange multipliers. Then,

$$P_\varepsilon = \inf_\gamma \sup_{\varphi,\psi} \mathcal{L}(\gamma,\varphi,\psi),$$

and the dual problem is obtained by interchanging the infimum and the supremum :

$$\begin{aligned} D_\varepsilon = \sup_{\varphi,\psi} \min_\gamma \sum_{x,y} \gamma_{xy}(c(x,y) - \psi(y) - \varphi(x) + \varepsilon(\log(\gamma_{xy}) - 1)) + \\ \sum_{x\in X} \varphi(x)\mu_x + \sum_{y\in Y} \psi(y)\nu_y. \end{aligned} \tag{1.6}$$

Taking the derivative with respect to γ_{xy}, we find that for a given φ, ψ, the optimal γ must satisfy:

$$\begin{aligned} c(x,y) - \psi(y) - \varphi(x) + \varepsilon \log(\gamma_{xy}) = 0 \\ \text{i.e. } \gamma_{xy} = \exp\left(\frac{\varphi(x) + \psi(y) - c(x,y)}{\varepsilon}\right) \end{aligned} \tag{1.7}$$

Putting these values in the definition of D_ε gives

$$\boxed{\begin{aligned} & D_\varepsilon = \sup_{\varphi,\psi} \Phi_\varepsilon(\varphi,\psi) \text{ with} \qquad (1.8) \\ & \Phi_\varepsilon(\varphi,\psi) := \sum_{x\in X} \varphi(x)\mu_x + \sum_{y\in Y} \psi(y)\nu_y - \sum_{x,y} \varepsilon \exp\left(\frac{\varphi(x)+\psi(y)-c(x,y)}{\varepsilon}\right) \end{aligned}}$$

Note that thanks to the relation (1.7), one can recover a solution to the primal problem from the dual one. This is true because, unlike the original linear programming formulation of the optimal transport problem, the regularized problem (1.2) is smooth and strictly convex. The following duality result holds

Theorem 1.3 (Strong duality). *Strong duality holds and the maximum in the dual problem is attained, that is $\exists \varphi, \psi$ such that*

$$P_\varepsilon = D_\varepsilon = \Phi_\varepsilon(\varphi, \psi).$$

Corollary 1.4. *If (φ, ψ) is the solution to (1.8), then the solution $\gamma^\star$ to (1.2) is given by*

$$\gamma_{x,y} = \exp\left(\frac{\varphi(x) + \psi(y) - c(x,y)}{\varepsilon}\right)$$

Notice now that the optimal coupling γ can be written as

$$\gamma_{x,y} = D_\varphi e^{\frac{-c(x,y)}{\varepsilon}} D_\psi,$$

where D_φ and D_ψ are the diagonal matrices associated to $e^{\varphi/\varepsilon}$ and $e^{\psi/\varepsilon}$, respectively. The problem is now similar to a matrix scaling problem

Definition 1.5 (Matrix scaling problem). Let $K \in \mathbb{R}^{N\times N}$ be a matrix with positive coefficients. Find D_ψ and D_ψ positive diagonal matrices in $K \in \mathbb{R}^{N\times N}$ such that $D_\varphi K D_\psi$ is doubly stochastic, that is sum along each row and each column is equal to 1.

Remark 1.6. Uniqueness fails since if (D_φ, D_ψ) is a solution then so is $(cD_\varphi, \frac{1}{c}D_\psi)$ for every $c \in \mathbb{R}_+$.

Algorithm 1 Sinkhorn-Knopp algorithm for the matrix scaling problem

```
function SINKHORN-KNOPP($K$)
    $D_\varphi^0 \leftarrow \mathbf{1}_N,\ D_\psi^0 \leftarrow \mathbf{1}_N$
    for $0 \leqslant k < k_{\max}$ do
        $D_\varphi^{k+1} \leftarrow \mathbf{1}_N./(KD_\psi^k)$
        $D_\psi^{k+1} \leftarrow \mathbf{1}_N./(K^T D_\varphi^{k+1})$
    end for
end function
```

Algorithm 2 Sinkhorn-Knopp algorithm for the regularised optimal transport problem

```
function SINKHORN-KNOPP($K_\varepsilon, \mu, \nu$)
    $D_\varphi^0 \leftarrow \mathbf{1}_X,\ D_\psi^0 \leftarrow \mathbf{1}_Y$
    for $0 \leqslant k < k_{\max}$ do
        $D_\varphi^{k+1} \leftarrow \mu./(KD_\psi^k)$
        $D_\psi^{k+1} \leftarrow \nu./(K^T D_\varphi^{k+1})$
    end for
end function
```

The matrix scaling problem can be easily solved by using an iterative algorithm, known as Sinkhorn-Knopp algorithm, which simply alternates updating D_φ and D_ψ in order to match the marginal constraints (a vector $\mathbf{1}_N$ of ones in this simple case).

where ./ stand for the element-wise division. Denoting by $(K_\varepsilon)_{x,y} = e^{\frac{-c(x,y)}{\varepsilon}}$ the algorithm takes the form 2 for the regularized optimal transport problem.

Notice that one can recast the regularized OT in the framework of bistochastic matrix scaling by replacing the kernel $e^{\frac{-c(x,y)}{\varepsilon}}$ with $(K_\varepsilon)_{x,y} = \mathrm{diag}(\mu)e^{\frac{-c(x,y)}{\varepsilon}}\,\mathrm{diag}(\nu)$, where $\mathrm{diag}(\mu)$ ($\mathrm{diag}(\nu)$) denotes the diagonal matrix with the vector μ (ν) as main diagonal. In this case the problem (1.2) can be re-written as

$$P_\varepsilon(\mu,\nu) = \inf\left\{ \langle\gamma|c\rangle + \varepsilon\mathcal{H}(\gamma|\mu\otimes\nu) \mid \gamma \in X\times Y,\ \sum_{y\in Y}\gamma_{xy} = \mu_x,\ \sum_{x\in X}\gamma_{xy} = \nu_y \right\}, \tag{1.9}$$

where $\mathcal{H}(\rho|\mu) := \sum_x \rho_x(\log(\frac{\rho_x}{\mu_x}) - 1)$ is the relative entropy or the Kullback-Leibler divergence.

Good to know: one can easily recast the regularized OT in the continuous framework as follows

$$\mathcal{P}_\varepsilon(\mu,\nu) = \inf\left\{ \int_{X\times Y} c(x,y)\mathrm{d}\gamma(x,y) + \varepsilon\mathcal{H}(\gamma|\mu\otimes\nu) \mid \gamma\in\Pi(\mu,\nu)\right\}, \tag{1.10}$$

where

$$\mathcal{H}(\rho|\pi) = \begin{cases} \int_{X\times Y}\Big(\log\Big(\frac{\mathrm{d}\rho(x,y)}{\mathrm{d}\pi(x,y)}\Big) - 1\Big)\mathrm{d}\rho(x,y), & \text{if } \rho \ll \pi \\ +\infty, & \text{otherwise,} \end{cases}$$

and the marginals μ,ν are probability measures on the compact metric spaces X and Y, respectively. This problem is often referred to as the *static Schrödinger problem* [9] since it was initially considered by Schrödinger in statistical physics. Once again, under mild assumptions on the cost functions, one can prove that the regularized problem converges to original one as $\varepsilon \to 0$; see [4, 8].

1.2 The convergence of Sinkhorn in the continuous setting

As presented in Lecture 1, the existence of Kantorovich potentials for the standard Optimal Transport problem can be proven by standard compactness arguments. By using similar arguments we show existence for the regularized dual problem (and convergence of Sinkhorn at the same time) in the continuous framework. We firstly recall that a coordinate ascent algorithm on a fucntion of two variables $f(x, y)$ can written as

$$\begin{aligned} y_{k+1} &= \operatorname{argmax}_y f(x_k, y), \\ x_{k+1} &= \operatorname{argmax}_x f(x, y_{k+1}). \end{aligned}$$

The Sinkhorn algorithm is actually a coordinate ascent algorithm: the main idea is indeed to maximize $\Phi_\varepsilon(\varphi, \psi)$ by maximizing alternatively in φ and ψ. From now on we assume for simplicity that $X = Y$ are compact and c is a continuous cost function.

Proposition 1.7. *The dual problem to* (1.10) *reads as*

$$D_\varepsilon = \sup\{\Phi_\varepsilon(\varphi, \psi) \mid \varphi, \psi \in \mathcal{C}_0(X)\}, \tag{1.11}$$

where

$$\begin{aligned} \Phi_\varepsilon(\varphi, \psi) := \int_X \varphi(x) \mathrm{d}\mu(x) &+ \int_Y \psi(y) \mathrm{d}\nu(y) \\ &- \varepsilon \int_{X \times Y} \exp\Big(\frac{\varphi(x) + \psi(y) - c(x, y)}{\varepsilon}\Big) \mathrm{d}\mu \otimes \mathrm{d}\nu(x, y). \end{aligned}$$

It is strictly concave w.r.t. each argument φ and ψ and strictly concave w.r.t. $\varphi(x) + \psi(y)$. It is also Fréchet differentiable for the $(\mathcal{C}_0, \|\cdot\|_\infty)$ topology. Furthermore, if a maximizer exists it is unique up to a constant, that is $\Phi_\varepsilon(\varphi, \psi) = \Phi_\varepsilon(\varphi + C, \psi - C)$ for every $C \in \mathbb{R}$.

Proof. We leave the proof as an exercice. □

Proposition 1.8. *The maximization of $\Phi_\varepsilon(\varphi, \psi)$ w.r.t. each variable can be made explicit, and the Sinkhorn algorithm can be defined as*

$$\varphi_{k+1}(x) = -\varepsilon \log\Big(\int_X \exp\Big(\frac{1}{\varepsilon}(\psi_k(y) - c(x, y)\Big) \mathrm{d}\nu(y)\Big) := S_\nu(\psi_k), \tag{1.12}$$

$$\psi_{k+1}(y) = -\varepsilon \log\Big(\int_X \exp\Big(\frac{1}{\varepsilon}(\varphi_{k+1}(x) - c(x, y)\Big) \mathrm{d}\mu(x)\Big) := S_\mu(\varphi_{k+1}). \tag{1.13}$$

Moreover, the following properties hold

(i) $\Phi_\varepsilon(\varphi_k, \psi_k) \leqslant \Phi_\varepsilon(\varphi_{k+1}, \psi_k) \leqslant \Phi_\varepsilon(\varphi_{k+1}, \psi_{k+1})$*;*

(ii) If $c(x, y)$ is $\omega-$continuous then $\varphi_{k+1}, \psi_{k+1}$ are also $\omega-$continuous ;

(iii) If $\psi_k - C$ $(\varphi_{k+1} - C)$ is bounded by M on the support of ν (μ), then so is φ_{k+1} (ψ_{k+1}).

Proof. (1.12) and (1.13) follow by writing the first-order necessary condition which gives us

$$1-\exp\left(\frac{\varphi(x)}{\varepsilon}\right)\int_Y \exp\Big(-\frac{1}{\varepsilon}(\psi(y)-c(x,y))\Big)\mathrm{d}\nu(y)=0,\ x-a.e.$$

implying the desired formula (and by symmetry, the same result on S_μ holds). Therefore, $S_\nu(\psi)$ is the unique maximizer of $\varphi \mapsto \Phi_\varepsilon(\varphi,\psi)$.
By definition of ascent on each coordinate, (i) is obtained directly. More generally one can prove that the application S_ν (S_μ) is $\omega-$continuous. Let $x_1,x_2\in X$ then

$$\begin{aligned}
&|S_\nu(\psi)(x_1)-S_\nu(\psi)(x_2)| = \varepsilon\log\left(\int_X e^{\left(\frac{1}{\varepsilon}(\psi(y)-c(x_2,y)\right)}\mathrm{d}\nu(y)\right)-\varepsilon\log\left(\int_X e^{\left(\frac{1}{\varepsilon}(\psi(y)-c(x_1,y)\right)}\mathrm{d}\nu(y)\right)\\
&=\varepsilon\log\left(\int_X e^{\left(\frac{1}{\varepsilon}(\psi(y)-c(x_1,y)+c(x_1,y)-c(x_2,y)\right)}\mathrm{d}\nu(y)\right)-\varepsilon\log\left(\int_X e^{\left(\frac{1}{\varepsilon}(\psi(y)-c(x_1,y)\right)}\mathrm{d}\nu(y)\right)\\
&\leqslant\varepsilon\log\left(e^{\frac{\omega(d(x_1,x_2))}{\varepsilon}}\int_X e^{\left(\frac{1}{\varepsilon}(\psi(y)-c(x_1,y)\right)}\mathrm{d}\nu(y)\right)-\varepsilon\log\left(\int_X e^{\left(\frac{1}{\varepsilon}(\psi(y)-c(x_1,y)\right)}\mathrm{d}\nu(y)\right)\\
&=\omega(d(x_1,x_2)).
\end{aligned}\tag{1.14}$$

The last point is just a bound on the iterates. □

Proposition 1.9. *The sequence* (φ_k,ψ_k) *defined by* (1.12) *and* (1.13) *converges in* $(\mathcal{C}_0,\|\cdot\|_\infty)$ *to the unique (up to a constant) couple of potentials* (φ,ψ) *which maximizes* Φ_ε.

Proof. Shifting the potentials by an additive constant, one can replace the optimization set by the couples (φ,ψ) which have uniformly bounded modulus of continuity and such that $\varphi(x_0)=0$ for a given $x_0\in X$. Recall that by proposition 1.7 the maximum of Φ is achieved at some couple (φ^*,ψ^*) which is unique up to a constant. Then, by prop. 1.8 (φ_k,ψ_k) are uniformly bounded and have uniformly modulus of continuity and one can extract a converging subsequence to $(\overline{\varphi},\overline{\psi})$. By continuity of Φ and the monotonicity of the sequence, $\Phi_\varepsilon(\overline{\varphi},S_\mu(\overline{\varphi}))\leqslant\Phi_\varepsilon(S_\nu\circ S_\mu(\overline{\varphi}),S_\mu(\overline{\varphi}))=\Phi_\varepsilon(\overline{\varphi},S_\mu(\overline{\varphi}))$. Therefore, the maximizer coordinatewise being unique, one has

$$S_\nu(\overline{\psi})=\overline{\varphi},\tag{1.15}$$

$$S_\mu(\overline{\varphi})=\overline{\psi}.\tag{1.16}$$

These show that $(\overline{\varphi},\overline{\psi})$ is a critical point for Φ_ε, thus being a maximizer. □

The proof of convergence relies on some important properties of the log−sum−exp (LSE) function $\log\int\exp$ which we summarise in the next Lemma. Before that let define the pseudo-norm $\|\cdot\|_{\circ,\infty}$ of uniform convergence as

$$\|f\|_{\circ,\infty}:=\frac{1}{2}(\sup f-\inf f)=\inf_{a\in\mathbb{R}}\|f+a\|_\infty\,.$$

Lemma 1.10. *The LSE function is convex and*

$$\|S_\mu(\varphi_1)-S_\mu(\varphi_2)\|_{\circ,\infty}\leqslant\|\varphi_1-\varphi_2\|_{\circ,\infty}\,.\tag{1.17}$$

Proof. Convexity is easily verified. We can get the 1−Lipschitz property as follows

$$\begin{aligned}|S_\mu(\varphi_1)(x) - S_\mu(\varphi_2)(x)| &= \left|\int_0^1 \frac{\mathrm{d}}{\mathrm{d}t} S_\mu(\varphi_2 + t(\varphi_1-\varphi_2))\mathrm{d}t\right| \\ &\leqslant \int_0^1 \left|\int_X (\varphi_1-\varphi_2)\frac{\exp(\frac{1}{\varepsilon}(\varphi_2 + t(\varphi_1-\varphi_2)-c))}{\int_X \exp(\frac{1}{\varepsilon}(\varphi_2+t(\varphi_1-\varphi_2)-c))\mathrm{d}\mu}\mathrm{d}\mu\right| \\ &\leqslant \|\varphi_1-\varphi_2\|_\infty .\end{aligned}$$

Notice that the equality occurs if and only if $\varphi_1 - \varphi_2$ is constant μ−a.e.. In particular we would have $\varphi_1 = \varphi_2 + a$ and $S_\mu(\varphi_1) = S_\mu(\varphi_2) + a$. Thus it is natural to consider the set of continuous functions up to an additive constant $\mathcal{C}(X)/\mathbb{R}$ endowed with the pseudo-norm introduced above. Then, since $S_\mu(\varphi_1 + a) = S_\mu(\varphi_1) + a$ we got the same inequality for the norm $\|\cdot\|_{\circ,\infty}$. □

Lemma 1.11. *Let $u, v \in \mathcal{C}(X)$ and $\mu \in \mathcal{P}(X)$ and denote ν_u and ν_v the Gibbs measures associated to u and v, that is $\mathrm{d}\nu_u = \frac{1}{Z_u}e^u\mathrm{d}\mu$ and $\mathrm{d}\nu_v = \frac{1}{Z_v}e^v\mathrm{d}\mu$, where Z_u and Z_v are the normalizing constants, then*

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

Proof. Consider a bounded function g on X and define

$$\eta_g(t) := \int_X g\frac{e^{tv+(1-t)u}}{Z_{t,g}}\mathrm{d}\mu,$$

where $Z_{t,g} = \int_X e^{tv+(1-t)u}\mathrm{d}\mu$. Differentiating we get

$$\eta_g'(t) + \eta_{v-u}(t)\eta_g(t) = \eta_{(v-u)g}(t),$$

and

$$e^{\int_0^t \eta_{v-u}(s)\mathrm{d}s}\eta_g(t) - \eta_g(0) = \int_0^t \eta_{(v-u)g}(s)e^{\int_0^s \eta_{v-u}(r)\mathrm{d}r}\mathrm{d}s.$$

Observe that

$$\begin{aligned}|e^{\int_0^t \eta_{v-u}(s)\mathrm{d}s}\eta_g(t) - \eta_g(0)| &\leqslant \|g\|_\infty \int_0^t \eta_{(u-v)}(s)e^{\int_0^s \eta_{u-v}(r)\mathrm{d}r}\mathrm{d}s \\ &\leqslant \|g\|_\infty\left(e^{\int_0^t \eta_{u-v}(s)\mathrm{d}s} - 1\right).\end{aligned}$$

Interchanging the role of u and v we have two possible cases: $\eta_g(1) \geqslant \eta_g(0) \geqslant 0$ or $\eta_g(1) \geqslant 0 \geqslant \eta_g(0)$. In the first case one has

$$|e^{\int_0^t \eta_{u-v}(s)\mathrm{d}s}(\eta_g(t) - \eta_g(0))| \leqslant |e^{\int_0^t \eta_{u-v}(s)\mathrm{d}s}\eta_g(t) - \eta_g(0)| \leqslant \|g\|_\infty\left(e^{\int_0^t \eta_{u-v}(s)\mathrm{d}s} - 1\right).$$

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

$$\begin{aligned}|\eta_g(1)| &\leqslant \|g\|_\infty \underbrace{\left(1 - e^{\int_{t_0}^1 \eta_{u-v}(s)\mathrm{d}s}\right)}_{:=a_1} \\ |\eta_g(0)| &\leqslant \|g\|_\infty \underbrace{\left(1 - e^{\int_0^{t_0} \eta_{u-v}(s)\mathrm{d}s}\right)}_{:=a_0}.\end{aligned}$$

Thus,

$$|\eta_g(1) - \eta_g(0)| \leqslant |\eta_g(1)| + |\eta_g(0)| \leqslant 2 \|g\|_\infty \max(a_1, a_0)$$

By exploiting the fact that $\eta_{u-v}(t) \leqslant 2 \|u - v\|_{\circ,\infty}$ we obtain in both cases that

$$\|\nu_u - \nu_v\| \leqslant 2(1 - e^{-2\|u-v\|_{\circ,\infty}})$$

□

Theorem 1.12. *(Convergence of Sinkhorn) The map $S = S_\nu \circ S_\mu$ is a contraction for $\|\cdot\|_{\circ,\infty}$. In particular the sequence (φ_k, ψ_k) defined by the Sinkhorn algorithm linearly converges to the unique (up to a constant) maximiser of the dual problem.*

Proof. We actually have to prove that

$$\|S_\mu(\varphi_1) - S_\mu(\varphi_2)\|_{\circ,\infty} \leqslant \kappa_\mu \|\varphi_1 - \varphi_2\|_{\circ,\infty}. \tag{1.18}$$

Once we have established that S_μ is a contraction then by lemma 1.10 it easily follows that

$$\|S(\varphi_1) - S(\varphi_2)\|_{\circ,\infty} \leqslant \kappa_\mu \|\varphi_1 - \varphi_2\|_{\circ,\infty},$$

which would conclude the proof.
In order to prove (1.18) we start by giving an estimation of the oscillations of S_μ

$$\frac{1}{2}|S_\mu(\varphi_1)(y) - S_\mu(\varphi_2)(y) - S_\mu(\varphi_1)(x) + S_\mu(\varphi_2)(x)| \leqslant \frac{1}{2}\left|\int_0^1 \int_X (\varphi_1 - \varphi_2)(\mathrm{d}\eta_{t,y} - \mathrm{d}\eta_{t,x})dt\right|,$$

where $\mathrm{d}\eta_{t,z} := \frac{1}{Z} e^{\frac{t(\varphi_1 - \varphi_2) + \varphi_2 - c(z,\cdot)}{\varepsilon}} \mathrm{d}\mu$ where Z is the normalising constant. Since $\mathrm{d}\eta_{t,z}$ is a Gibbs measure we can apply the L^1 bound of lemma 1.11 to estimate $\|\eta_{t,y} - \eta_{t,x}\|_{L^1}$ and get

$$\|S_\mu(\varphi_1) - S_\mu(\varphi_2)\|_{\circ,\infty} \leqslant \kappa_\mu \|\varphi_1 - \varphi_2\|_{\circ,\infty}$$

with $\kappa_\mu = (1 - e^{-2\frac{\|c\|_{\circ,\infty}}{\varepsilon}})$. □

Remark 1.13 (Convergence speed)**.** This theorem shows that the Sinkhorn algorithm converges linearly, but notice that the contraction constant has a bad dependency in ε. Denoting $C = \|c\|_{\circ,\infty}$, to get an error of β one needs

$$(1 - e^{-2\frac{C}{\varepsilon}})^k \leqslant \beta$$

that is

$$k \gtrapprox e^{2C/\varepsilon} \log(1/\beta).$$

Remark 1.14. We refer the interested reader to [3, 10] where the convergence of Sinkhorn algorithm in infinite dimension (and generalized also to the multi-marginal case) is treated.

A The convergence of Sinkhorn for the Hilbert metric in the discrete setting

We focus now on the global convergence analysis of the Sinkhorn algorithm in the discrete setting by using the *Hilbert* projective metric on $\mathbb{R}^n_{+,\star}$ (positive vectors).

Definition A.1 (Hilbert projective metric)**.** The *Hilbert* projective metric on $\mathbb{R}^n_{+,\star}$ is defined as

$$\forall (u,v) \in (\mathbb{R}^n_{+,\star})^2,\ d_H(u,v) := ||\log(u) - \log(v)||_V,$$

Where

$$||x||_V = \max_i x_i - \min_i x_i.$$

Before stating the convergence result we need the following fundamental theorem, which shows that a positive matrix is a strict contraction on the cone of positive vector

Theorem A.2 ([2, 12])**.** *Let $K \in \mathbb{R}^{n\times n}_{+,\star}$, then for $(u,v) \in (\mathbb{R}^n_{+,\star})^2$*

$$d_H(Ku, Kv) \leqslant \lambda(K) d_H(u,v), \tag{A.19}$$

where

$$\lambda(K) = \frac{\sqrt{\eta(K)} - 1}{\sqrt{\eta(K)} + 1} < 1$$

and

$$\eta(K) = \max_{i,j,kl} \frac{K_{ik}K_{jl}}{K_{jk}K_{il}}.$$

We have then the following convergence result (we use the same notations as in 2)

Theorem A.3 ([7])**.** *One has $(D^k_\varphi, D^k_\psi) \to (D^\star_\varphi, D^\star_\psi)$ and*

$$d_H(D^k_\varphi, D^\star_\varphi) = O(\lambda(K)^{2k}),\ d_H(D^k_\psi, D^\star_\psi) = O(\lambda(K)^{2k}), \tag{A.20}$$

where $D^\star_\varphi, D^\star_\psi$ are the optimal solutions. Moreover,

$$d_H(D^k_\varphi, D^\star_\varphi) \leqslant \frac{d_H(\gamma^k \mathbf{1}_n, \mu)}{1 - \lambda(K)^2}, \tag{A.21}$$

$$d_H(D^k_\psi, D^\star_\psi) \leqslant \frac{d_H(\gamma^k \mathbf{1}_n, \nu)}{1 - \lambda(K)^2}, \tag{A.22}$$

where $\gamma^k = \operatorname{diag}(D^k_\varphi) K \operatorname{diag}(D^k_\psi)$. Last, one has

$$||\log(\gamma^k) - \log(\gamma^\star)||_\infty \leqslant d_H(D^k_\varphi, D^\star_\varphi) + d_H(D^k_\psi, D^\star_\psi). \tag{A.23}$$

where $\gamma^\star$ is the unique solution to (1.2).

Proof. Notice that for any $(u,v) \in (\mathbb{R}^n_{+,\star})^2$, on has

$$d_H(u,v) = d_H(u/v, \mathbf{1}_n) = d_H(\mathbf{1}_n/u, \mathbf{1}_n/v).$$

This shows that

$$d_H(D^k_\varphi, D^\star_\varphi) = d_H(\frac{\mu}{KD^k_\psi}, \frac{\mu}{KD^\star_\psi}) = d_H(KD^k_\psi, KD^\star_\psi) \leqslant \lambda(K) d_H(D^k_\psi, D^\star_\psi),$$

where we used Theorem A.2. This shows (A.20). By using triangular inequality we have

$$\begin{aligned} d_H(D^k_\varphi, D^\star_\varphi) &\leqslant d_H(D^{k+1}_\varphi, D^k_\varphi) + d_H(D^{k+1}_\varphi, D^\star_\varphi) \\ &\leqslant d_H(\frac{\mu}{KD^k_\psi}, D^k_\varphi) + \lambda(K) d_H(D^k_\varphi, D^\star_\varphi) \\ &= d_H(\mu, D^k_\varphi \odot (KD^k_\psi)) + \lambda(K)^2 d_H(D^k_\varphi, D^\star_\varphi) \\ &= d_H(\mu, \gamma^k \mathbf{1}_n) + \lambda(K)^2 d_H(D^k_\varphi, D^\star_\varphi), \end{aligned}$$

where $\odot$ denotes the element wise multiplication. (A.22) can be proved in an analogous way. (A.23) is trivial. □

References

[1] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré, *Iterative bregman projections for regularized transportation problems*, SIAM Journal on Scientific Computing **37** (2015), no. 2, A1111–A1138.

[2] Garrett Birkhoff, *Extensions of jentzsch's theorem*, Transactions of the American Mathematical Society **85** (1957), no. 1, 219–227.

[3] Guillaume Carlier, *On the linear convergence of the multi-marginal sinkhorn algorithm*, (2021).

[4] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer, *Convergence of entropic schemes for optimal transport and gradient flows*, SIAM Journal on Mathematical Analysis **49** (2017), no. 2, 1385–1418.

[5] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard, *Scaling algorithms for unbalanced transport problems*, arXiv preprint arXiv:1607.05816 (2016).

[6] Marco Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, Advances in neural information processing systems, 2013, pp. 2292–2300.

[7] Joel Franklin and Jens Lorenz, *On the scaling of multidimensional matrices*, Linear Algebra and its applications **114** (1989), 717–735.

[8] Christian Léonard, *From the schr\" odinger problem to the monge-kantorovich problem*, arXiv preprint arXiv:1011.2564 (2010).

[9] ———, *A survey of the schr\" odinger problem and some of its connections with optimal transport*, arXiv preprint arXiv:1308.0215 (2013).

[10] Simone Di Marino and Augusto Gerolin, *An optimal transport approach for the schrödinger bridge problem and convergence of sinkhorn algorithm*, Journal of Scientific Computing **85** (2020), no. 2.

[11] Gabriel Peyré, Marco Cuturi, et al., *Computational optimal transport*, Foundations and Trends® in Machine Learning **11** (2019), no. 5-6, 355–607.

[12] Hans Samelson et al., *On the perron-frobenius theorem.*, The Michigan Mathematical Journal **4** (1957), no. 1, 57–59.

[13] François-Xavier Vialard, *An elementary introduction to entropic regularization and proximal methods for numerical optimal transport*, (2019).