Entropic regularization for optimal transport

Bernhard Schmitzer

Summer school optimal transport, DFG SPP 1962, Dortmund, September 2023

Disclaimer

- material selected here is based on personal experience with the subject so far
- no longer able to keep track of all developments; lots of interesting work on the limit $\varepsilon \to 0$
- need to skip many interesting aspects, also some technical details
- goal is not that all details here will be understood at first pass; important: now get the big picture, can follow up on details on second pass through notes

1 Introduction

Definition 1.1 (Setting)**.**

- X will be a compact metric space, $C(X)$ denotes continuous (real-valued) functions
- $\mathcal{M}(X)$, $\mathcal{M}_+(X)$, $\mathcal{P}(X)$ will be (signed) Radon measures, non-negative measures, and probability measures, respectively
- We will use a lot the duality between continuous functions and measures on X.
- $c \in C(X \times X)$ will be a continuous cost function
- P_1, P_2 denote the marginal projection operators:

$$\int_X \phi(x) \mathrm{d}(P_1\gamma)(x) := \int_{X\times X} \phi(x) \mathrm{d}\gamma(x,y)$$

  correspond to push-forward by map $(x_1, x_2) \mapsto x_i$.
- For finite spaces $X = \{x_1, \dots, x_n\}$, we usually identify $\mathcal{M}(X) \simeq C(X) \simeq \mathbb{R}^n$ (and likewise with product spaces). Then for $\mu \in \mathcal{M}(X)$ we denote by μ_i the mass at x_i, et cetera.

Definition 1.2 (KL divergence)**.** Let

$$\varphi : \mathbb{R} \to [0,\infty], \qquad s \mapsto \begin{cases} s\log(s) - s + 1 & \text{for } s > 0, \\ 1 & \text{for } s = 0, \\ +\infty & \text{for } s < 0. \end{cases}$$

Note that φ is convex, proper and lower-semicontinuous. Then for $\mu, \nu \in \mathcal{M}(X)$, the Kullback–Leibler (KL) divergence of μ w.r.t. ν is given by

$$\mathrm{KL}(\mu|\nu) := \begin{cases} \int_X \varphi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\nu & \text{if } \mu \ll \nu,\ \nu \geq 0, \\ +\infty & \text{else.} \end{cases}$$

Definition 1.3 (Entropic transport problem)**.** For $\mu, \nu \in \mathcal{P}(X)$, a continuous cost function $c \in C(X \times X)$, regularization strength $\varepsilon > 0$ and a reference measure $\rho \in \mathcal{M}_+(X \times X)$, the entropic transport problem is given by

$$C_\varepsilon(\mu,\nu) := \inf \left\{ \int_{X\times X} c \,\mathrm{d}\gamma + \varepsilon\ \mathrm{KL}(\gamma|\rho) \middle| \gamma \in \Gamma(\mu,\nu) \right\}$$

Remark 1.4 (Motivation)**.**

- Will show: entropic transport problem has unique solution; gives stability of solutions w.r.t. fluctuations in marginals or cost function; even differentiability. Useful for downstream applications.
- Can be solved efficiently with simple numerical method: Sinkhorn algorithm
- Will allow for more reliable statistical estimation of optimal transport between sampled empirical measures. (Beyond the scope of this course.)

Remark 1.5 (Choice of reference measure)**.** There are many different potential choices for the reference measure $\rho \in \mathcal{M}_+(X \times X)$. Common choices are:

(i) For a finite space $X = \{x_1, \dots, x_n\}$ one often uses the Shannon entropy, i.e. one sets ρ to be the counting measure, $\rho_{i,j} = 1$.

(ii) In some applications, if $X \subset \mathbb{R}^d$, it may be natural to choose $\rho = \mathcal{L}^{2d} \llcorner (X \times X)$. This will only work well, if all measures of interest μ and ν are dominated by the Lebesgue measure.

(iii) The most agnostic choice is probably $\rho = \mu \otimes \nu$. Then, the optimal objective is always finite (for bounded cost functions) and we will show that regular dual solutions exist.

Lemma 1.6. KL is jointly weak* lower-semicontinuous and convex in both arguments. For fixed $\nu \in \mathcal{M}_+(X)$, $\mu \mapsto \mathrm{KL}(\mu|\nu)$ is strictly convex.

Proof. Joint lsc follows from [Ambrosio *et al.*, 2000, Theorem 2.34] since the function φ is non-negative, convex and lower-semicontinuous. Join convexity follows from the fact that the function $(r, s) \mapsto s \log(s/r) - s + r$. Strict convexity in the first argument follows from strict convexity of φ. □

Lemma 1.7. If the infimal objective is finite, the entropic optimal transport problem has a unique solution.

Proof. Since $\Gamma(\mu, \nu)$ is bounded (and weak* closed), any minimizing sequence has weak* cluster points that also lie in $\Gamma(\mu, \nu)$. Clearly, the map $\gamma \mapsto \int_{X \times X} c \mathrm{d}\gamma$ is continuous. By the above Lemma the KL term is lower-semicontinuous. Hence, any cluster point must be a minimizer.
Assume the objective is finite. Let $\gamma_1, \gamma_2 \in \Gamma(\mu, \nu)$ be two minimizers, and therefore have $\mathrm{KL}(\gamma_i|\rho) < \infty$. By linearity of the term $\gamma \mapsto \int_{X \times X} c \mathrm{d}\gamma$ and strict convexity of the KL-term, if $\gamma_1 \neq \gamma_2$, then $(\gamma_1 + \gamma_2)/2$ would be an even better candidate. Hence $\gamma_1 = \gamma_2$ and the minimizer must be unique. □

Remark 1.8. Choose $c(x, y) = d(x, y)^p$ for a metric d on X and $\varepsilon > 0$. Then the map

$$W_\varepsilon : \mathcal{P}(X)^2 \ni (\mu, \nu) \mapsto C_\varepsilon(\mu, \nu)^{1/p}$$

is no longer a metric on $\mathcal{P}(X)$. (Unless something really boring happens, like X being only a single point and choosing ρ to be the probability measure on that single point.)
In general one finds that $W_\varepsilon(\mu, \mu) > 0$ and that it violates the triangle inequality. The former issue (along with something called 'entropic bias') can be fixed by the Sinkhorn divergence [Feydy *et al.*, 2018]. While they do not satisfy the triangle inequality, they are a useful (and statistically robust) notion of similarity in many applications. The question of a metric induced by entropic optimal transport is (to my knowledge) still open.

2 Convergence as $\varepsilon \to 0$

Definition 2.1 (Setting)**.** Let $(\varepsilon_n)_{n \in \mathbb{N}}$ be a strictly positive, decreasing sequence with limit $\lim_{n \to \infty} \varepsilon_n = 0$. Let $\mu, \nu \in \mathcal{P}(X)$ and choose as reference measure $\rho = \mu \otimes \nu$. Set

$$E_n(\gamma) := \int_{X \times X} c \, \mathrm{d}\gamma + \varepsilon_n \, \mathrm{KL}(\gamma|\mu \otimes \nu) + \iota_{\Gamma(\mu,\nu)}(\gamma),$$
$$E(\gamma) := \int_{X \times X} c \, \mathrm{d}\gamma + \iota_{\Gamma(\mu,\nu)}(\gamma).$$

Let $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence of minimizers of E_n (existence shown above, with finite optimal objective). We will now show that, up to selection of subsequences, $(\gamma_n)_n$ converges to some γ that minimizes E, and that the optimal objectives converge.

Proposition 2.2 (Finite space)**.** Let $X = \{x_1, \dots, x_n\}$ be a finite space. Then the sequence $(\gamma_n)_{n\in\mathbb{N}}$ is precompact and any cluster point γ minimizes E. The optimal objectives converge.

Proof. For simplicity, w.l.o.g. assume that μ and ν are strictly positive (otherwise, simply remove points x_i where $\mu_i = 0$ from the first marginal space, and similarly with the second marginal). Then ρ is strictly positive, and by finiteness of X, ρ is bounded away from zero by some finite constant. Therefore, by continuity of φ on its domain $[0, \infty]$, the domain of

$$\mathbb{R}^{n\times n} \ni \gamma \mapsto \mathrm{KL}(\gamma|\rho) = \sum_{i,j} \varphi(\gamma_{i,j}/\rho_{i,j}) \cdot \rho_{i,j}$$

is $\mathbb{R}_+^{n\times n}$ and it is continuous on this domain. The set $\Gamma(\mu,\nu) \subset \mathbb{R}_+^{n\times n}$ is compact and therefore $\mathrm{KL}(\cdot|\rho)$ is bounded on $\Gamma(\mu,\nu)$ (say, by some constant $C < \infty$). Therefore, $E_n(\gamma_n) - E(\gamma_n) \in [0, \varepsilon_n \cdot C]$.
Since $\Gamma(\mu,\nu)$ is compact and non-empty, $(\gamma_n)_n$ will have cluster points (that also lie in $\Gamma(\mu,\nu)$). Let now γ be any such cluster point, for simplicity denote by $(\gamma_n)_n$ a convergent subsequence. Clearly, E is continuous on $\Gamma(\mu,\nu)$. Therefore we find:

$$\lim_n E_n(\gamma_n) = \lim_n E(\gamma_n) = E(\gamma)$$

Finally, if there were some $\tilde{\gamma} \in \Gamma(\mu,\nu)$ with $E(\tilde{\gamma}) < E(\gamma)$, then for sufficiently big n (and some suitable $\delta > 0$) we would have

$$E_n(\tilde{\gamma}) = E(\tilde{\gamma}) + \varepsilon_n \cdot \mathrm{KL}(\tilde{\gamma}|\mu\otimes\nu) \le E(\tilde{\gamma}) + \varepsilon_n \cdot C < E(\gamma) - \delta \le E(\gamma_n) \le E_n(\gamma_n)$$

which contradicts the optimality of γ_n for E_n. Therefore, γ must minimize E. □

For continuous X the situation is more involved, since $\mathrm{KL}(\cdot|\mu\otimes\nu)$ will in general not be bounded (or even finite) on $\Gamma(\mu,\nu)$ and might be infinite for minimizers of E. We will show convergence of minimizers by means of Γ-convergence. Much more general versions of the following result are possible (on non-compact spaces, with less regular cost functions, and with approximate marginals). We focus on a few key properties of the problem here.

Lemma 2.3. Let $(\gamma_n)_n$ be a sequence in $\mathcal{M}(X \times X)$ that converges weak* to $\gamma \in \mathcal{M}(X \times X)$. Then

$$\liminf_n E_n(\gamma_n) \ge E(\gamma).$$

Proof. Since $\Gamma(\mu,\nu)$ is weak* closed, if $\gamma \notin \Gamma(\mu,\nu)$, we will eventually have that $\gamma_n \notin \Gamma(\mu,\nu)$ and thus $E(\gamma) = E_n(\gamma_n) = \infty$ for sufficiently large n. So assume $\gamma \in \Gamma(\mu,\nu)$ from now on.
Then, using weak* continuity of the linear cost term and non-negativity of the entropy term, we obtain

$$\begin{aligned}\liminf_n E_n(\gamma_n) &= \liminf_n \int_{X\times X} c\,\mathrm{d}\gamma_n + \varepsilon_n \cdot \mathrm{KL}(\gamma_n|\mu\otimes\nu) \\ &\ge \lim_n \int_{X\times X} c\,\mathrm{d}\gamma_n = E(\gamma).\end{aligned}$$

□

The lim-sup inequality is considerably more involved. For $\gamma \in \Gamma(\mu,\nu)$ with $\mathrm{KL}(\gamma|\mu\otimes\nu) = \infty$ we need to construct an approximating sequence $(\gamma_n)_n$ with finite entropy (diverging in a controlled way) while preserving the marginals. We do this here via the *block approximation trick* [Carlier *et al.*, 2017].

Definition 2.4 (Block approximation)**.** Let $\gamma \in \Gamma(\mu,\nu)$. For a length scale $L > 0$, denote by $\{X_{L,i}\}_{i=1}^{n_L}$ a (measurable) partition of X into n_L sets, each of which with diameter at most L. Such a partition exists by compactness of X. Denote in the following

$$\mu_{L,i} := \mu(X_{L,i}), \qquad \nu_{L,i} := \nu(X_{L,i}), \qquad \gamma_{L,i,j} := \gamma(X_{L,i} \times X_{L,j}),$$

and finally

$$\lambda_{L,i,j} := \begin{cases} \frac{\mu \llcorner X_{L,i} \otimes \nu \llcorner X_{L,j}}{\mu_{L,i} \cdot \nu_{L,j}} & \text{if } \mu_{L,i} \cdot \nu_{L,j} > 0, \\ 0 & \text{else.} \end{cases}$$

Then the *block approximation* of γ at scale L is given by

$$\gamma_L := \sum_{i,j=1}^{n_L} \gamma_{L,i,j} \cdot \lambda_{L,i,j}.$$

Lemma 2.5. $\gamma_L \in \Gamma(\mu,\nu)$.

Proof. First observe: $\gamma_L \geq 0$. Next, observe that

$$\sum_{j=1}^{n_L} \gamma_{L,i,j} = \sum_{j=1}^{n_L} \gamma(X_{L,i} \times X_{L,j}) = \gamma(X_{L,i} \times X) = \mu(X_{L,i}) = \mu_{L,i}.$$

In particular, this implies $\gamma_{L,i,j} > 0 \Rightarrow \mu_{L,i} > 0$. And of course likewise for the other marginal. Now, for any measurable $A \subset X$ one has

$$\begin{aligned} \gamma_L(A \times X) &= \sum_{i,j=1}^{n_L} \gamma_{L,i,j} \cdot \lambda_{L,i,j}(A \times X) \\ &= \sum_{\substack{i,j=1,\dots,n_L: \\ \gamma_{L,i,j}>0}} \frac{\gamma_{L,i,j}}{\mu_{L,i} \cdot \nu_{L,j}} \cdot \mu(X_{L,i} \cap A) \cdot \nu(X_{L,j} \cap X) \\ &= \sum_{\substack{i,j=1,\dots,n_L: \\ \gamma_{L,i,j}>0}} \frac{\gamma_{L,i,j}}{\mu_{L,i}} \cdot \mu(X_{L,i} \cap A) \\ &= \sum_{\substack{i=1,\dots,n_L: \\ \mu_{L,i}>0}} \mu(X_{L,i} \cap A) = \mu(A). \end{aligned}$$

The same computation applies for the second marginal, which completes the proof. □

Lemma 2.6. Equip $X \times X$ with the metric $D((x_1,x_2),(y_1,y_2)) := d(x_1,y_1) + d(x_2,y_2)$ (which yields a compact metric space). Then $W_p(\gamma,\gamma_L) \leq 2L$ and in particular $\gamma_L \overset{*}{\rightharpoonup} \gamma$ as $L \to 0$.

Proof. A potential transport plan from γ to γ_L involves moving mass only within products of partition cells $X_{L,i} \times X_{L,j}$, which have diameter bounded by $2L$ in D. This yields the Wasserstein bound, which implies the weak* convergence. □

Lemma 2.7. $\mathrm{KL}(\gamma_L | \mu \otimes \nu) \leq 2\log(n_L)$.

Proof.

$$\begin{aligned}
\mathrm{KL}(\gamma_L|\mu\otimes\nu) &= \int_{X\times X} \varphi\left(\tfrac{\mathrm{d}\gamma_L}{\mathrm{d}\mu\otimes\nu}\right)\mathrm{d}\mu\otimes\nu \\
&= \sum_{\substack{i=1,\dots,n_L:\\ \mu_{L,i}>0}} \sum_{\substack{j=1,\dots,n_L:\\ \nu_{L,j}>0}} \varphi\left(\frac{\gamma_{L,i,j}}{\mu_{L,i}\cdot\nu_{L,j}}\right)\cdot\mu_{L,i}\cdot\nu_{L,j} \\
&= \sum_{\substack{i=1,\dots,n_L:\\ \mu_{L,i}>0}} \sum_{\substack{j=1,\dots,n_L:\\ \nu_{L,j}>0}} \left[\gamma_{L,i,j}\cdot\log\left(\frac{\gamma_{L,i,j}}{\mu_{L,i}\cdot\nu_{L,j}}\right) - \gamma_{L,i,j} + \mu_{L,i}\cdot\nu_{L,j}\right] \\
&\leq -\sum_{\substack{i=1,\dots,n_L:\\ \mu_{L,i}>0}} \sum_{\substack{j=1,\dots,n_L:\\ \nu_{L,j}>0}} \gamma_{L,i,j}\cdot[\log(\mu_{L,i})+\log(\nu_{L,j})] \\
&\leq -\sum_{\substack{i=1,\dots,n_L:\\ \mu_{L,i}>0}} \mu_{L,i}\cdot\log(\mu_{L,i}) - \sum_{\substack{j=1,\dots,n_L:\\ \nu_{L,j}>0}} \nu_{L,j}\cdot\log(\nu_{L,j}) \\
&\leq -2\log(1/n_L) = 2\log(n_L)
\end{aligned}$$

where we used that $\mathbb{R}_+^{n_L} \ni p \mapsto \sum_{i=1}^{n_L} p_i\log(p_i)$ is convex and minimized among 'probability vectors' by the 'uniform' one $p_i = 1/n_L$. □

Using now the block approximation and its basic properties that we established, we conclude:

Lemma 2.8 (Lim sup)**.** For any $\gamma\in\Gamma(\mu,\nu)$ there is a sequence $(\gamma_n)_n$, converging weak* to γ, such that

$$\lim_n E_n(\gamma_n) = E(\gamma).$$

Proof. Let $(L_n)_n$ be a positive sequence. Then

$$E_n(\gamma_n) \leq \int_{X\times X} c\,\mathrm{d}\gamma_n + 2\varepsilon_n\log(n_{L_n}).$$

Choosing now $(L_n)_n$ decreasing, such that $\varepsilon_n\log(n_{L_n})\to 0$, and with the weak* convergence of $(\gamma_{L_n})_n$ to γ one obtains the result. □

Together, lim-inf and lim-sup inequality provide:

Proposition 2.9. Let $(\gamma_n)_n$ be a sequence of minimizers for E_n (their existence was established earlier). Then the sequence is weak* precompact and any cluster point minimizes E.

Proof. Weak* precompactness is obtained from compactness of $\Gamma(\mu,\nu)$. Any cluster point γ then satisfies

$$\liminf_n E_n(\gamma_n) \geq E(\gamma).$$

(Restriction to subsequences is no issue here, since $E_n(\gamma_n)$ can be seen to be non-increasing, and thus all subsequences have the same limit inferior.) If there were some other $\tilde\gamma$, with $E(\tilde\gamma) < E(\gamma)$, then a recovery sequence $(\tilde\gamma_n)_n$ for $\tilde\gamma$, constructed via the block approximation as above, would satisfy

$$E(\tilde\gamma) = \lim_n E_n(\tilde\gamma_n)$$

and thus for sufficiently big n one would obtain the contradiction

$$E_n(\tilde\gamma_n) < E_n(\gamma_n).$$

□

3 Duality

Theorem 3.1 (Fenchel–Rockafellar)**.** Let (X,X^*), (Y,Y^*) be two couples of topologically paired spaces. Let $A: X\to Y$ be a bounded linear operator. Let G and F be proper convex functions,

defined on X and Y respectively, with values in $(-\infty, \infty]$. If there exists $x \in X$ such that G is finite at x and F is continuous at Ax, then

$$\inf_{x \in X} F(Ax) + G(x) = \sup_{y^* \in Y^*} -F^*(-y^*) - G^*(A^* y^*).$$

The supremum is attained.

Proposition 3.2 (Duality for entropic OT)**.** A dual problem for the entropic OT problem is given by

$$C_\varepsilon(\mu, \nu) = \sup \left\{ \int_X \phi \, \mathrm{d}\mu + \int_X \psi \, \mathrm{d}\nu - \varepsilon \int_{X \times X} \left[\exp\left(\frac{\phi \oplus \psi - c}{\varepsilon} \right) - 1 \right] \mathrm{d}\rho \middle| \phi, \psi \in C(X) \right\}.$$

Here $\phi \oplus \psi$ denotes the function $(x, y) \mapsto \phi(x) + \psi(y)$.

Proof. Setting

$$\begin{aligned} &F : \mathcal{M}(X)^2 \to (-\infty, \infty], && F = \iota_{\{(\mu,\nu)\}}, \\ &G : \mathcal{M}(X \times X) \to (-\infty, \infty], && \gamma \mapsto \int_{X \times X} c \, \mathrm{d}\gamma + \varepsilon \, \mathrm{KL}(\gamma | \rho), \\ &A : \mathcal{M}(X \times X) \to \mathcal{M}(X)^2, && \gamma \mapsto (P_1 \gamma, P_2 \gamma) \end{aligned}$$

we can write the entropic transport problem as

$$\inf_{\gamma \in \mathcal{M}(X \times X)} F(A\gamma) + G(\gamma).$$

We find that we cannot directly apply the FR duality theorem, since F is nowhere continuous. We will still proceed for now and observe in the end, that we can indeed apply the theorem in the 'reverse' direction by flipping the roles of primal and dual problem (and keeping careful track of minus signs).
We find:

$$F^*((\phi, \psi)) = \sup_{(\rho,\sigma) \in \mathcal{M}(X)^2} \int_X \phi \, \mathrm{d}\rho + \int_X \psi \, \mathrm{d}\sigma - \iota_{\{(\mu,\nu)\}}(\rho, \sigma) = \int_X \phi \, \mathrm{d}\mu + \int_X \psi \, \mathrm{d}\nu.$$

For the adjoint of A:

$$\langle (\phi, \psi), A\gamma \rangle = \int_X \phi \mathrm{d} P_1 \gamma + \int_X \psi \mathrm{d} P_2 \gamma = \int_{X \times X} [\phi(x) + \psi(y)] \mathrm{d}\gamma(x, y)$$

and so $A^*(\phi, \psi) = \phi \oplus \psi$. For G we first observe that $G(\gamma) = \int_{X \times X} c \, \mathrm{d}\gamma + \varepsilon \, \mathrm{KL}(\gamma | \mu \otimes \nu)$. That is, it is obtained from KL first by a positive re-scaling, and then by adding a linear term. Using the simple relations

$$[f(x) = \varepsilon \cdot g(x)] \quad \Rightarrow \quad [f^*(z) = \varepsilon \cdot g^*(z/\varepsilon)]$$

$$[f(x) = \langle a, x \rangle + g(x)] \quad \Rightarrow \quad [f^*(z) = g^*(z - a)]$$

we obtain that $G^*(\xi) = \varepsilon \, \mathrm{KL}^*((\xi - c)/\varepsilon | \rho)$ where KL^* denotes the conjugation with respect to the first argument. One obtains that

$$\begin{aligned} \mathrm{KL}^*(\xi | \rho) &= \sup_{\gamma \in \mathcal{M}(X \times X)} \int \xi \, \mathrm{d}\gamma - \mathrm{KL}(\gamma | \rho) \\ &= \sup_{u \in L^1(\rho)} \int [\xi \cdot u - \varphi(u)] \, \mathrm{d}\rho = \int \varphi^*(\xi) \, \mathrm{d}\rho \end{aligned}$$

and a brief explicit computation yields that $\varphi^*(s) = \exp(s) - 1$. Then, formally writing down

$$\sup_{(\phi,\psi) \in C(X)^2} -F^*(-(\phi, \psi)) - G^*(A^*(\phi, \psi))$$

yields the above expression for the dual problem.
It remains to show that we can actually apply the FR theorem. For this we now observe that the functions F^* and G^* are globally finite and continuous, hence the 'reverse constraint qualifications' are satisfied. This implies that FR also provides the existence of optimal entropic OT plans, which we had already established earlier by direct methods. □

Proposition 3.3 (Primal-dual optimality conditions for Fenchel–Rockafellar duality)**.** x and $y*$ are primal and dual optimal in the FR-primal-dual problem pair above if and only if

$$[Ax \in \partial F^*(-y^*) \Leftrightarrow -y^* \in \partial F(Ax)] \wedge [x \in \partial G^*(A^*y^*) \Leftrightarrow A^*y^* \in \partial G(x)].$$

Proof. The Fenchel–Young inequality states that

$$F(x) + F^*(y*) \geq \langle x, y* \rangle$$

with equality if and only if $x \in \partial F^*(y^*)$ or equivalently $y^* \in \partial F(x)$.
Now consider the primal dual gap of the above problem pair:

$$\begin{aligned} 0 &\leq [F(Ax) + G(x)] - [-F^*(-y^*) - G^*(A^*y)] \\ &= [F(Ax) + F^*(-y^*) + \langle Ax, -y^* \rangle] + [G(x) + G^*(A^*y) + \langle x, A^*y^* \rangle] \end{aligned}$$

By the Fenchel–Young inequality, this can be zero if and only if both parentheses are zero, which happens if and only if the subdifferential conditions for both apply, which are the stated PD optimality conditions. □

Proposition 3.4 (Application to entropic OT)**.** A pair $\gamma \in \mathcal{M}(X \times X)$, $(\phi, \psi) \in C(X)^2$ are primal-dual optimal if and only if

$$P_1\gamma = \mu, \qquad P_2\gamma = \nu, \qquad \gamma = \exp([\phi \oplus \psi - c]/\varepsilon) \cdot \rho.$$

Proof. Consider the condition $A\gamma \in \partial F^*(-(\phi, \psi))$. Since F^* is the linear pairing $(\alpha, \beta) \mapsto \langle \alpha, \mu \rangle + \langle \beta, \nu \rangle$, it is subdifferential is the singleton (μ, ν) at all points. With A being the marginal projection operator, this translates to the two marginal constraints for γ.
The function $G^*(\xi) = \varepsilon \int [\exp([\xi - c]/\varepsilon) - 1] \, \mathrm{d}\rho$ is differentiable with

$$\frac{\mathrm{d}}{\mathrm{d}t} G^*(\xi + t \cdot \eta)|_{t=0} = \int \exp([\xi - c]/\varepsilon) \eta \mathrm{d}\rho.$$

The subdifferential is therefore given by $\partial G^*(\xi) = \exp([\xi - c]/\varepsilon)\rho$. Inserting now the argument $\xi = \phi \oplus \psi$ yields the expression for γ. □

4 Sinkhorn algorithm

Remark 4.1 (Choice of reference measure and existence of optimal dual solutions)**.** For this section we fix the choice $\rho = \mu \otimes \nu$. This will have some useful consequences, in particular existence of continuous (and even more regular of c is 'nice') optimal dual solutions.

The Sinkhorn algorithm can be interpreted as alternating block optimization on the dual problem, alternatingly fixing one of the functions ϕ or ψ and optimizing over the other.

Lemma 4.2. For fixed $\psi \in C(X)$, an optimal ϕ in the dual entropic OT problem is given by

$$\phi(x) = -\varepsilon \cdot \log \left(\int_X \exp([\psi(y) - c(x, y)]/\varepsilon) \, \mathrm{d}\nu(y) \right).$$

Of course, the corresponding Lemma for the second marginal also holds.

Proof. The dual entropic OT objective is concave and differentiable. Hence, its maximizers can be determined by studying the first order optimality conditions. Let

$$J(\phi, \psi) := \int \phi \, \mathrm{d}\mu + \int \psi \mathrm{d}\nu - \varepsilon \int_{X \times X} [\exp([\phi \oplus \psi - c]/\varepsilon) - 1] \, \mathrm{d}\mu \otimes \nu.$$

One finds that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} J(\phi + t \cdot \eta, \psi)|_{t=0} &= \int_X \eta \, \mathrm{d}\mu - \int_{X \times X} \exp([\phi(x) + \psi(y) - c(x, y)]/\varepsilon) \eta(x) \mathrm{d}\mu(x) \mathrm{d}\nu(y) \\ &= \int_X \left[1 - \exp(\phi(x)/\varepsilon) \int_X \exp([\psi(y) - c(x, y)]/\varepsilon) \mathrm{d}\nu(y) \right] \eta(x) \, \mathrm{d}\mu(x) \end{aligned}$$

For this to be zero for all $\eta \in C(X)$, we need that the bracket is zero μ-almost everywhere. Resolving this expression for ϕ yields the given expression. □

Definition 4.3 (Sinkhorn algorithm)**.** For some initial $\psi^{(0)} \in C(X)$, set recursively for $\ell \in \{0, 1, \dots\}$,

$$\phi^{(\ell+1)}(x) = -\varepsilon \cdot \log\left(\int_X \exp([\psi^{(\ell)}(y) - c(x,y)]/\varepsilon)\,\mathrm{d}\nu(y)\right),$$
$$\psi^{(\ell+1)}(y) = -\varepsilon \cdot \log\left(\int_X \exp([\phi^{(\ell+1)}(x) - c(x,y)]/\varepsilon)\,\mathrm{d}\mu(x)\right).$$

We refer to this procedure as the Sinkhorn algorithm.

Remark 4.4 (Primal interpretation of dual optimality condition)**.** Recall the primal-dual optimality condition $\gamma = \exp([\phi \oplus \psi - c]/\varepsilon) \cdot \mu \otimes \nu$ and the marginal constraint $P_1\gamma = \mu$. Together these imply for all $\eta \in C(X)$,

$$\begin{aligned}\int_X \eta(x)\,\mathrm{d}\mu(x) &= \int_{X\times X} \eta(x)\,\mathrm{d}\gamma(x,y)\\ &= \int_X \left[\exp(\phi(x)/\varepsilon)\int_X \exp([\psi(y) - c(x,y)]/\varepsilon)\mathrm{d}\nu(y)\right]\eta(x)\,\mathrm{d}\mu(x).\end{aligned}$$

This is precisely the dual optimality condition for ϕ obtained in the above lemma. Hence, this dual optimality condition has the primal interpretation that ϕ is chosen just right, such that the implied primal γ has the prescribed first marginal μ. Alternating maximization therefore also has the interpretation of alternating re-scaling. Consequently, the Sinkhorn algorithm is also known as iterative proportional fitting procedure.
Finally, these re-scalings can also be interpreted as KL projections of the 'old' γ onto one of the two marginal constraints, and thus the algorithm can also be interpreted as alternating projection method.

Remark 4.5 (Entropic c-transform)**.** The map $\psi \mapsto \psi^{c,\nu,\varepsilon} := \phi$ where ϕ is given by the locally optimal ϕ,

$$\phi(x) = -\varepsilon \cdot \log\left(\int_X \exp([\psi(y) - c(x,y)]/\varepsilon)\,\mathrm{d}\nu(y)\right)$$

is sometimes referred to as entropic c-transform.
The 'classical' c-transform is given by

$$\psi^c(x) = \inf_{y\in X} c(x,y) - \psi(y)$$

and it plays an important role in the analysis of unregularized optimal transport and some numerical methods, such as the auction algorithm.
Applying the classical c-transform in unregularized optimal transport will however always become stationary after at most three steps ($\phi^{ccc} = \phi^c$) and need not be converge to dual maximizers. This is related to the fact that the unregularized dual problem is constrained and therefore non-smooth. Alternating maximization may therefore get stuck in the 'ridge' corresponding to the constraint.

Similar to the classical c-transform, $\psi^{c,\nu,\varepsilon}$ inherits some regularity from c.

Lemma 4.6. Assume that c has a modulus of continuity ω in its first argument, i.e. $c(x+\delta, y) \le c(x,y) + \omega(|\delta|)$ (here $\omega : [0,\infty) \to [0,\infty)$ is continuous and $\omega(0) = 0$), then $\psi^{c,\nu,\varepsilon}$ has the same modulus of continuity.

Proof.

$$\begin{aligned}\psi^{c,\nu,\varepsilon}(x+\delta) &= -\varepsilon \cdot \log\left(\int_X \exp([\psi(y) - c(x+\delta,y)]/\varepsilon)\,\mathrm{d}\nu(y)\right)\\ &\le -\varepsilon \cdot \log\left(\int_X \exp([\psi(y) - c(x,y) - \omega(|\delta|)]/\varepsilon)\,\mathrm{d}\nu(y)\right)\\ &\le \psi^{c,\nu,\varepsilon}(x) + \omega(|\delta|).\end{aligned}$$

□

Remark 4.7. Indeed, one can show even higher-order Sobolev type regularity of entropic c-transforms for the squared distance cost, see for instance [Genevay *et al.*, 2019]. This is a crucial step for statistical stability of empirical (entropic) optimal transport. This higher-order regularity deteriorates as $\varepsilon \to 0$ and in the limit one obtains the 'cursed' convergence rates of classical optimal transport.

This can be used for a simple convergence proof of the Sinkhorn algorithm, based on compactness.

Proposition 4.8. The Sinkhorn algorithm converges (up to subsequences and optimal constant shifts) to a solution of the dual entropic OT problem. In particular, optimal dual solutions exist.

Proof. Since $c \in C(X \times X)$, there exists a modulus of continuity for both arguments. Hence, the Sinkhorn iterates all have the same modulus of continuity and they are therefore equi-continuous. Adding a constant shift to each $\phi^{(\ell+1)}$ such that $\phi^{(\ell+1)}(x_0) = 0$ for some arbitrary fixed $x_0 \in X$ will merely result in a corresponding shift in the opposite direction in $\psi^{(\ell+1)}$. In particular the sequence $(\phi^{(\ell)})_\ell$ will also be equibounded, as will be the sequence $(\psi^{(\ell)})_\ell$.
So by the Arzela–Ascoli theorem, there exists a pair of cluster points (ϕ, ψ) such that a suitable subsequence of iterates (with some constant shifts) converges uniformly to these two functions. Since the entropic c-transform is continuous, this means that $\phi = \psi^{c,\nu,\varepsilon}$ and $\psi = \phi^{c^\top,\mu,\varepsilon}$ (here $c^\top$ denotes the 'flipped' cost function with first and second argument flipped).
This implies that $\gamma := \exp([\phi \oplus \psi - c]/\varepsilon) \cdot \mu \otimes \nu$ satisfies both marginal constraints, and thus the tripled $(\gamma, (\phi, \psi))$ is primal and dual optimal for the entropic OT problem. □

Remark 4.9.

- The convergence (and optimality proof) above does not extend to the case $\varepsilon = 0$, since being a fixed point of the c-transforms is not sufficient for optimality in the unregularized problem.
- The choice $\rho = \mu \otimes \nu$ was important in this section, as only by this choice does the expression for $\psi^{c,\nu,\varepsilon}$ obtain spatial regularity independent of μ. This yields the compactness of the iterates and existence of a continuous dual solutions.

Remark 4.10 (Matrix-scaling formulation)**.** Introduce the functions

$$u := \exp(\phi/\varepsilon), \qquad\qquad v := \exp(\psi/\varepsilon),$$

(and likewise applied to all the iterates of the Sinkhorn algorithm). Then the iterations can be written as

$$u^{(\ell+1)}(x) = 1/\int \exp(-c(x,y)/\varepsilon) v^{(\ell)}(y)\, \mathrm{d}\nu(y),$$
$$v^{(\ell+1)}(y) = 1/\int \exp(-c(x,y)/\varepsilon) u^{(\ell+1)}(x)\, \mathrm{d}\mu(x).$$

This can easily be expressed as matrix-vector multiplications in the discrete setting. This might be a tad faster than the logsumexp-version, but is also numerically more prone to issues, especially for small ε. There are many tricks for running the Sinkhorn algorithm stable and efficiently at small ε.

Remark 4.11 (Speed of convergence)**.**

- [Franklin and Lorenz, 1989]: linear convergence of dual iterates to maximizer in Hilbert's projective metric. But: contraction ratio approaches 1 like $1 - \exp(-\|c\|_\infty/\varepsilon)$ as $\varepsilon \to 0$.
- [Schmitzer, 2019]: convergence of an asymmetric ('auction-like') Sinkhorn algorithm in $O(1/\varepsilon)$ iterations (measured in L^1-error of primal iterate marginal constraints)
- [Berman, 2020]: convergence of the Sinkhorn algorithm for the W_2 distance on the Torus in $O(1/\varepsilon)$ iterations, by showing that the iterates asymptotically follow a non-linear PDE
- ε-scaling very efficient in practice (at least on 'normal problems') but no proof for its efficiency yet (as far as I am aware).

- There are several variants of Sinkhorn, intended to be faster, such as the 'Greenkhorn' algorithm.

Remark 4.12 (Flexibility of the Sinkhorn algorithm)**.** One of the biggest strengths of the Sinkhorn algorithm is that it can easily be adapted to related problems, such as optimal transport barycenters, multi-marginal transport problems (only efficient, if there is some trick to handle the high problem dimensionality), and unbalanced transport problems. See for instance: [Benamou *et al.*, 2015; Peyré, 2015; Chizat *et al.*, 2018; Benamou *et al.*, 2019]

References

L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems*. Oxford mathematical monographs. Oxford University Press, 2000.

Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM J. Imaging Sci.*, 37(2):A1111–A1138, 2015.

Jean-David Benamou, Guillaume Carlier, and Luca Nenna. Generalized incompressible flows, multi-marginal transport and Sinkhorn algorithm. *Numerische Mathematik*, 142(1):33–54, 2019.

Robert J. Berman. The Sinkhorn algorithm, parabolic optimal transport and geometric Monge–Ampère equations. *Numerische Mathematik*, 145:771–836, 2020.

Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM J. Math. Anal.*, 49(2):1385–1418, 2017.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Math. Comp.*, 87:2563–2609, 2018.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. arXiv:1810.08278, 2018.

Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114–115:717–735, 1989.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 1574–1583, 2019.

Gabriel Peyré. Entropic approximation of Wasserstein gradient flows. *SIAM J. Imaging Sci.*, 8(4):2323–2351, 2015.

Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM J. Sci. Comput.*, 41(3):A1443–A1481, 2019.