Entropic regularization of Optimal Transport

Quentin Bertrand, Mathurin Massias, Titouan Vayer

Last updated: December 4, 2024

Contents

1 Entropic regularization of optimal transport

References: Peyré et al. (2019, Chap. 4).

Definition 1.1 (Entropy)**.** *The entropy of a positive vector $x \in \mathbb{R}_+^d$ is:*

$$H(x) = -\sum_{i=1}^{d} x_i \log(x_i) \tag{1.1}$$

with the convention $0 \log 0 = 0$. It extends to matrices by summing over both indices i and j: $H(P) = -\sum_{i,j} P_{ij} \log(P_{ij})$.

The Hessian of the entropy is $\mathrm{diag}(-1/x_i)$ *so the entropy is* ***strictly concave****. In 1D on $[0, 1]$ it looks roughly like a concave parabola, as shown on Figure 1, but with infinite slope at 0.*

The negative entropy *is simply minus the entropy. It is a negative quantity, and strictly convex.*

Exercise 1.1. *Show that the element of the simplex that minimizes the negative entropy (equivalently, that maximizes the entropy) is the constant vector $(1/d, \dots, 1/d)$.*

Definition 1.2. *The entropic-regularized optimal transport problem (EOT) is:*

$$\min_{P \geq 0} \langle C, P \rangle + \varepsilon \sum_{ij} P_{ij}(\log(P_{ij}) - 1) \quad s.t. \quad P1 = a,\ P^\top 1 = b \tag{1.2}$$

Since the negative entropy is strictly convex, the solution of EOT is unique. Note that the -1 in the entropy is here for convenience and can be removed without affecting the solution, since the sum of entries of P is constant for any feasible P.

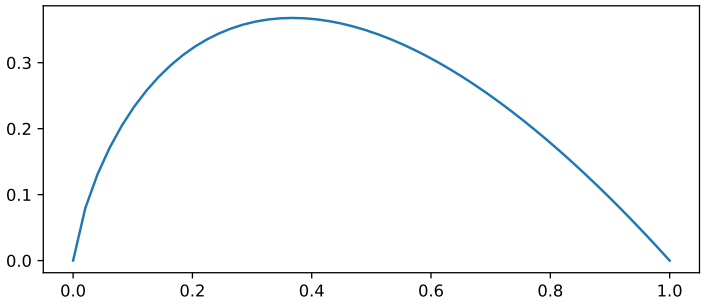

Figure 1: the entropy

Proposition 1.3. *When ε goes to 0 the solution P^* of the entropic problem converges to a solution of the Kantorovich problem (which corresponds to $\varepsilon = 0$). If there are multiple solutions to the Kantorovich problem, it converges to the one that has maximal entropy.*

Note that this result is not true for general optimization problems: argmin and limit do not commute.

Proof. Take a sequence of positive regularization strength ε_ℓ going to 0, with associated EOT solutions P_ℓ. Since the constraint set is the same for all P_ℓ, and that it is closed and bounded, we can extract a converging subsequence (say towards P_∞) that we rename P_ℓ for convenience. Let P^* be a solution of the Kantorovich problem. Then, by optimality of P^* for the Kantorovich problem, since P_ℓ is feasible for it too,

$$\langle C, P* \rangle \leq \langle C, P_\ell \rangle \tag{1.3}$$

On the other hand, since P^* is feasible for EOT with $\varepsilon = \varepsilon_\ell$, by optimality of P_ℓ for this problem,

$$\langle C, P_\ell \rangle - \varepsilon_\ell H(P_\ell) \leq \langle C, P^* \rangle - \varepsilon_\ell H(P^*) \tag{1.4}$$

Combining the two we get:

$$0 \leq \langle C, P_\ell \rangle - \langle C, P^* \rangle \leq \varepsilon_\ell (H(P_\ell) - H(P^*)) \tag{1.5}$$

Since (P_ℓ) is bounded (the feasible set is the same for all problem, and it is bounded), so is $H(P_\ell)$; by dividing by ε_ℓ and letting ℓ to infinity we must have $\langle C, P_\infty \rangle = \langle C, P^* \rangle$, which shows that P_∞ is a solution of the Kantorovich problem (as it is feasible). □

Exercise 1.2. *Show that when $\varepsilon \to \infty$, the solution of EOT goes to the feasible point that has maximal entropy. Show that this point is $ab^\top$.*

Definition 1.4. *The* Gibbs kernel *associated to the EOT problem is $K = \exp(-\frac{C}{\varepsilon})$, where the exponential acts entrywise. It is thus a matrix with strictly positive entries.*

Definition 1.5. *The Kullback-Leibler divergence between two matrices $P, Q \in \mathbb{R}_+^{m\times n}$ (so, with positive entries) is:*

$$\mathrm{KL}(P,Q) \triangleq \sum_{i=1}^{m}\sum_{j=1}^{n} P_{ij}\log\frac{P_{ij}}{Q_{ij}} - Q_{ij} + P_{ij} \tag{1.6}$$

Proposition 1.6. *The KL divergence is the Bregman divergence associated to the negative entropy. It is thus strictly convex in its first variable.*

Proposition 1.7 (EOT as Bregman projection)**.** *The solution of EOT is also the solution of:*

$$\operatorname*{argmin}_{P} \mathrm{KL}(P,K) \quad s.t. \quad P \geq 0,\ P1 = a,\ P^\top 1 = b \tag{1.7}$$

with K the Gibbs kernel $K = \exp(-C/\varepsilon)$.

This means that solving EOT is the same as projecting the Gibbs kernel onto the feasible set, in the sense of the KL divergence (and not, as in the usual Euclidean projection, in the sense of $\frac{1}{2}\|\cdot - \cdot\|^2$).

Proposition 1.8. *The dual of EOT is the following unconstrained problem:*

$$\max_{f,g} \langle a, f\rangle + \langle b, g\rangle - \varepsilon\sum_{ij}\exp\left(\frac{f_i + g_j - C_{ij}}{\varepsilon}\right) \tag{1.8}$$

We see that as $\varepsilon \to 0$, the values for which $f_i + g_j > C_{ij}$ are more and more penalized, eventually leading to a hard constraint for $\varepsilon = 0$.

In addition, for any solution f^*, g^* to the dual, the primal solution P^* satisfies:

$$P^*_{ij} = \exp\left(\frac{f^*_i + g^*_j - C_{ij}}{\varepsilon}\right) \tag{1.9}$$

One of the strong benefits of EOT is that its dual is unconstrained, so when solving it one does not have to enforce constraints such as $f_i + g_j \leq C_{ij}$, which are nearly impossible to apply in the continuous case.

Exercise 1.3. *Show that it (f^*, g^*) is a solution to the dual of EOT, so is $(f^* + \delta 1_m, g^* - \delta 1_n)$ for any value of δ.*

Fun fact: show that if a and b have entries all smaller than 1, then any solution (f^*, g^*) of the dual problem must satisfy $f^*_i + g^*_j \leq C_{ij}$ – and this, even though the constraint is not enforced in a hard way, unlike in unregularized OT.

Definition 1.9. *Let $M \in \mathbb{R}_{++}^{m\times n}$, $a \in \mathbb{R}^m_{++}$, and $b \in \mathbb{R}^n_{++}$. The* matrix scaling problem *is to find vectors $u, v \in \mathbb{R}^m_{++} \times \mathbb{R}^n_{++}$ such that*

$$\mathrm{diag}(u) M\,\mathrm{diag}(v) 1 = a \tag{1.10}$$

$$(\mathrm{diag}(u) M\,\mathrm{diag}(v))^\top 1 = b \tag{1.11}$$

It corresponds to multiplying rows and columns of M by scalars so that the rows end up summing to a and the columns end up summing to b.

Exercise 1.4. *Show that if u, v is a solution of the matrix scaling problem, so is $(u/\lambda, \lambda v)$ for any $\lambda > 0$.*

Proposition 1.10. *The matrix scaling problem with $M = K$ the Gibbs kernel is equivalent to the dual of EOT: if (u^*, v^*) solves the matrix scaling problem, then it is equal to $(\exp(f^*/\varepsilon), \exp(g^*/\varepsilon))$ where (f^*, g^*) solves the dual of EOT (and vice versa, any solution of the dual can be mapped to a solution of the matrix scaling problem).*

How do the two types of invariance (by addition-subtraction for f, g, by multiplication/division for (u, v)) relate to this?

Definition 1.11. *The Sinkhorn algorithm starts from a strictly positive v^0 and iterates:*

$$u^{\ell+1} = a/Kv^{\ell} \tag{1.12}$$

$$v^{\ell+1} = b/K^{\top}u^{\ell+1} \tag{1.13}$$

where division is meant pointwise.

Sinkhorn's algorithm has many interpretation. The easiest one if the following:

- for a fixed v, the choice of u that makes $\operatorname{diag}(u)K\operatorname{diag}(v)$ satisfy $\operatorname{diag}(u)K\operatorname{diag}(v)1 = a$ is a/Kv (this is because $K\operatorname{diag}(v)1 = Kv$).
- for a fixed u, the choice of v that makes $\operatorname{diag}(u)K\operatorname{diag}(v)$ satisfy $(\operatorname{diag}(u)M\operatorname{diag}(v))^{\top}1 = b$ is $b/K^{\top}u$.

So, Sinkhorn alternatively modifies u and v to satisfy the first and the second constraints.

Surprisingly, this works: Sinkhorn's algorithm converges to a solution of the matrix scaling problem, and hence can be used to solve the dual of EOT.

Proposition 1.12. *If we define $P^{\ell} = \operatorname{diag}(u^{\ell})K\operatorname{diag}(v^{\ell})$, then the sequence P^{ℓ} corresponds to alternated Bregman projection (in the KL sense) onto the constraint set.*

Alternating projections usually do not converge, but when the constraint sets are affine, it does.

Proposition 1.13. *If one defines $u = \exp(f/\varepsilon)$ and $v = \exp(g/\varepsilon)$, then performing iterations of Sinkhorn is the same as performing alternate maximization (in f and in g) on the dual of EOT.*

1.1 Convergence of the Sinkhorn algorithm

To derive convergence results easily we'll work in a different metric

Definition 1.14. *The* Hilbert projective metric *on $\mathbb{R}^d_{++}$ is:*

$$d_{\mathcal{H}}(u, u) \triangleq \log \max_{i,j} \frac{u_i u'_j}{u_j u'_i} \tag{1.14}$$

Interpretation: it is a distance on the cone $\mathbb{R}^d_{++}$ quotiented by the equivalence relation $u \sim u' \Leftrightarrow u = ru', r > 0$.

This metric is useful because it "does not see" variations in scaling: $d_{\mathcal{H}}(u, \lambda u') = d_{\mathcal{H}}(u, u')$; those variations in scaling in u and v are precisely the one that makes the solutions to the matrix scaling problem not unique.

Exercise 1.5. *Show that the Hilbert projective metric satisfies the triangular inequality* $d_{\mathcal{H}}(u, u'') \leq d_{\mathcal{H}}(u, u') + d_{\mathcal{H}}(u', u'')$.

Proposition 1.15. *Let* $M \in \mathbb{R}_{++}^{m \times n}$. *Then* M *is a contraction in the Hilbert projective metric: for any positive* v, v',

$$d_{\mathcal{H}}(Mv, Mv') \leq \lambda(M) d_{\mathcal{H}}(v, v') \tag{1.15}$$

where $\lambda(K) \triangleq \frac{\sqrt{\eta(M)}-1}{\sqrt{\eta(M)}+1} < 1$ *and* $\eta(M) \triangleq \max_{i,j,k,l} \frac{M_{ik} M_{kl}}{M_{jk} M_{il}} \geq 1$ *is some sort of "conditioning" of* K.

Proposition 1.16. *Consider the iterates of Sinkhorn algorithm* u^ℓ, v^ℓ, *and define* $P^\ell = \operatorname{diag}(u^\ell) K \operatorname{diag}(v^\ell)$. *Then*

1. $u^\ell, v^\ell \to u^*, v^*$ *a solution of the matrix scaling problem associated to* K, *and so* P^ℓ *converges to the solution of EOT. The convergence is at a speed* $d_{\mathcal{H}}(u^\ell, u^*) = \mathcal{O}(\lambda(K)^{2\ell})$, *same for* v.
2. $d_{\mathcal{H}}(u^\ell, u^*) \leq \frac{d_{\mathcal{H}}(P^\ell 1, a)}{1-\lambda(K)^2}$ *and similarly for* v *and* b.
3. $\|P^\ell - P^*\|_\infty \leq d_{\mathcal{H}}(u^\ell, u^*) + d_{\mathcal{H}}(v^\ell, v^*)$.

References

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 2019.