Screening Sinkhorn Algorithm via Dual Projections

Abstract

Computing optimal transport distances, such as the earth mover's distance is a fundamental problem in machine learning, statistics, and computer vision.

1 Introduction

Related work.

Our contribution.

Notation. We denote Σ_n the probability simplex with n bins, namely the set of probability vectors in $\mathbb{R}_+^n$, i.e., $\Sigma_n = \{w \in \mathbb{R}_+^n : \sum_{i=1}^n w_i = 1\}$. For any positive matrix $T \in \mathbb{R}^{n \times m}$, we define its negative entropy as $H(T) = -\sum_{i,j} T_{ij} \log(T_{ij})$. Let $r(T) = T\mathbf{1}_m \in \mathbb{R}^n$ and $c(T) = T^\top \mathbf{1}_n \in \mathbb{R}^m$ denote the rows and columns sums of T respectively. The coordinates $r_i(T)$ and $c_j(T)$ denote the i-th row sum and the j-th column sum of T, respectively. The scalar product between two matrices denotes the usual inner product, that is $\langle T, W \rangle = \text{tr}(T^\top W) = \sum_{i,j} T_{ij} W_{ij}$, where $T^\top$ is the transpose of T. We write $\mathbf{1}$ (resp. $\mathbf{0}$) the vector having all coordinates equal to one (resp. zero). $\Delta(w)$ denotes the diag operator, such that if $w \in \mathbb{R}^n$, then $\Delta(w) = \text{diag}(w_1, \ldots, w_n) \in \mathbb{R}^{n \times n}$. Throughout this paper, when applied to matrices and vectors, $\odot$ and $\oslash$ (Hadamard product and division) and exponential notations refer to elementwise operators. We also denote $|I|$ the cardinality of a finite set I. Given two real numbers a and b, we write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$.

2 Regularized optimal transport

We briefly present in this section the setup of optimal transport between two discrete measures. We then consider the case when those distributions are only available through a finite number of samples, that is $\mu = \sum_{i=1}^n \mu_i \delta_{x_i} \in \Sigma_n$ and $\nu = \sum_{j=1}^m \nu_i \delta_{y_j} \in \Sigma_m$. We denote their probabilistic couplings set as $\Pi(\mu, \nu) = \{P \in \mathbb{R}_+^{n \times m}, P\mathbf{1}_m = \mu, P^\top \mathbf{1}_n = \nu\}$.

Sinkhorn divergence. Approximating the optimal transport distance between the two measures μ and ν amounts to solving a linear problem given by [3]

$$\mathcal{S}(\mu, \nu) = \min_{P \in \Pi(\mu, \nu)} \langle M, P \rangle, \tag{1}$$

where $P = (P_{ij}) \in \mathbb{R}^{n \times m}$ is called the transportation plan, namely each entry P_{ij} represents the fraction of mass moving from x_i to y_j, and $M = (M_{ij}) \in \mathbb{R}^{n \times m}$ is a cost matrix comprised of nonnegative elements related to the energy needed to move a probability mass from x_i to y_j. The entropic regularization of optimal transport distances [2] relies on the addition of a penalty term as follows:

$$\mathcal{S}_\eta(\mu, \nu) = \min_{P \in \Pi(\mu, \nu)} \{\langle M, P \rangle - \eta H(P)\}, \tag{2}$$

where $\eta > 0$ is a regularization parameter. We refer to $\mathcal{S}_\eta(\mu, \nu)$ as the *Sinkhorn divergence* [2].

Preprint. Under review.

Dual of Sinkhorn divergence. Below we provide the derivation of the dual problem for the regularized optimal transport problem (2). Towards this end, we begin with writing its Lagrangian dual function :

$$\mathscr{L}(P, y, z) = \langle M, P\rangle + \eta\langle \log P, P\rangle + \langle y, P\mathbf{1}_m - \mu\rangle + \langle z, P^\top\mathbf{1}_n - \nu\rangle,$$

which can be rewritten in the following form:

$$\mathscr{L}(P, y, z) = -\langle y, \mu\rangle - \langle z, \nu\rangle + \langle P, M + \eta \log P\rangle + \langle y, P\mathbf{1}_m\rangle + \langle z, P^\top\mathbf{1}_n\rangle.$$

The dual of Sinkhorn divergence can be derived by solving $\min_{P\in\mathbb{R}_+^{n\times m}} \mathscr{L}(P, y, z)$. It is easy to check that objective function $P \mapsto \mathscr{L}(P, y, z)$ is strongly convex and differentiable. Hence, one can solve the latter minimum by setting $\nabla_P\mathscr{L}(P, y, z)$ to $\mathbf{0}_{n\times m}$. Therefore, we get

$$P^\star_{ij} = \exp\Big(-\frac{1}{\eta}(y_i + z_j + M_{ij}) - 1\Big), \tag{3}$$

for all $i = 1, \ldots, n$ and $j = 1, \ldots, m$. Plugging this solution, one has

$$\min_{P\in\mathbb{R}_+^{n\times m}} \mathscr{L}(P, y, z) = -\langle y, \mu\rangle - \langle z, \nu\rangle - \eta\sum_{i,j}\exp\Big(-\frac{1}{\eta}(y_i + z_j + M_{ij}) - 1\Big).$$

Setting the change of variables $u = -y/\eta - 1/2$ and $v = -z/\eta - 1/2$, we get

$$\min_{P\in\mathbb{R}_+^{n\times m}} \mathscr{L}(P, y, z) = \eta\Big(\langle u, \mu\rangle + \langle v, \nu\rangle - \sum_{i,j}\exp\big(u_i + v_j - M_{ij}/\eta)\big) + 1\Big).$$

Recall that the dual problem is given by $\max_{y\in\mathbb{R}^n, z\in\mathbb{R}^m}\min_{P\in\mathbb{R}_+^{n\times m}} \mathscr{L}(P, y, z)$, that is

$$\max_{u\in\mathbb{R}^n, v\in\mathbb{R}^m}\Big\{\eta\Big(\langle u, \mu\rangle + \langle v, \nu\rangle - \sum_{i,j}\exp\big(u_i + v_j - M_{ij}/\eta)\big) + 1\Big)\Big\},$$

which is equivalent to solving

$$\max_{u\in\mathbb{R}^n, v\in\mathbb{R}^m}\Big\{\langle u, \mu\rangle + \langle v, \nu\rangle - \sum_{i,j}\exp\big(u_i + v_j - M_{ij}/\eta)\big)\Big\}$$

and which can be rewritten in the following matrix form:

$$\min_{u\in\mathbb{R}^n, v\in\mathbb{R}^m}\big\{\Psi(u, v) := \mathbf{1}_n^\top B(u, v)\mathbf{1}_m - \langle u, \mu\rangle - \langle v, \nu\rangle\big\}, \tag{4}$$

where $B(u, v) = \Delta(e^u)K\Delta(e^v)$ and $K = e^{-M/\eta}$ stands for the Gibbs kernel associated to the cost matrix M. We refer to problem (4) to the *dual of Sinkhorn divergence*. Therefore, the optimal solution of Sinkhorn divergence is given by $P^\star = \Delta(e^{u^\star})K\Delta(e^{v^\star})$ where the couple $(u^\star, v^\star)$ satisfies:

$$(u^\star, v^\star) = \operatorname*{argmin}_{u\in\mathbb{R}^n, v\in\mathbb{R}^m}\{\Psi(u, v)\}.$$

Note that the matrices $\Delta(e^{u^\star})$ and $\Delta(e^{v^\star})$ are unique up to a constant factor [4].

3 Screened dual of Sinkhorn divergence

For a fixed $\varepsilon > 0$ and $\kappa > 0$ we define an *approximate dual of Sinkhorn divergence*

$$\min_{u\in\mathcal{C}^n_{\frac{\varepsilon}{\kappa}}, v\in\mathcal{C}^m_{\varepsilon\kappa}}\big\{\Psi_\kappa(u, v) := \mathbf{1}_n^\top B(u, v)\mathbf{1}_m - \langle \kappa u, \mu\rangle - \langle \kappa^{-1}v, \nu\rangle\big\}, \tag{5}$$

where $\mathcal{C}^r_\alpha \subseteq \mathbb{R}^r$, for $r \in \mathbb{N}$ and $\alpha > 0$, is a convex set given by $\mathcal{C}^r_\alpha = \{w \in \mathbb{R}^r : \min_{1\le i\le r} e^{w_i} \ge \alpha\}$.

The κ-parameter in problem (5) plays a role of scaling factor, namely it allows to get a closed order of the potential vectors e^u and e^v, while the ε-parameter acts like a threshold for e^u and e^v. Note that the setting of $\varepsilon = 0$ and $\kappa = 1$, the approximate dual of Sinkhorn divergence coincides with the dual of Sinkhorn divergence (4).

The screening procedure presented in this work is based on constructing two *active sets* $I_{\varepsilon,\kappa}$ and $J_{\varepsilon,\kappa}$ throughout the dual problem of (5) in the following way:

Lemma 1. *Let* (u^*, v^*) *be an optimal solution of the problem* (5)*. Define*

$$I_{\varepsilon,\kappa} = \Big\{i = 1, \ldots, n : \mu_i \geq \varepsilon^2\kappa^{-1} r_i(K)\Big\}, J_{\varepsilon,\kappa} = \Big\{j = 1, \ldots, m : \nu_j \geq \kappa\varepsilon^2 c_j(K)\Big\}$$

and $I^{\complement}_{\varepsilon,\kappa} = \{1, \ldots, n\}\backslash I_{\varepsilon,\kappa}$*, and* $J^{\complement}_{\varepsilon,\kappa} = \{1, \ldots, n\}\backslash J_{\varepsilon,\kappa}$*. Then one has* $e^{u_i^*} = \varepsilon\kappa^{-1}$ *and* $e^{v_j^*} = \varepsilon\kappa$ *for all* $i \in I^{\complement}_{\varepsilon,\kappa}$ *and* $j \in J^{\complement}_{\varepsilon,\kappa}$*.*

Proof. Introducing two dual variables $\lambda \in \mathbb{R}^n_+$ and $\beta \in \mathbb{R}^m_+$ for each constraint, the Lagrangian of problem (5) reads as

$$\mathscr{L}(u, v, \lambda, \beta) = \frac{\varepsilon}{\kappa}\langle\lambda, \mathbf{1}_n\rangle + \varepsilon\kappa\langle\beta, \mathbf{1}_m\rangle + \mathbf{1}_n^\top B(u, v)\mathbf{1}_m - \langle\kappa u, \mu\rangle - \langle\kappa^{-1}v, \nu\rangle - \langle\lambda, e^u\rangle - \langle\beta, e^v\rangle$$

First order conditions [1] then yield that the Lagrangian multiplicators solutions λ^* and β^* satisfy the following:

$$\nabla_u\mathscr{L}(u^*, v^*, \lambda^*, \beta^*) = e^{u^*} \odot (Ke^{v^*} - \lambda^*) - \kappa\mu = \mathbf{0}_n,$$
$$\text{and } \nabla_v\mathscr{L}(u^*, v^*, \lambda^*, \beta^*) = e^{v^*} \odot (K^\top e^{u^*} - \beta) - \kappa^{-1}\nu = \mathbf{0}_m$$

which leads to

$$\lambda^* = Ke^{v^*} - \kappa\mu \oslash e^{u^*} \text{ and } \beta^* = K^\top e^{u^*} - \nu \oslash \kappa e^{v^*}$$

For all $i = 1, \ldots, n$ we have that $e^{u_i^*} \geq \frac{\varepsilon}{\kappa}$. By the KKT optimality conditions, the condition on the dual variable $\lambda_i^* > 0$ ensures that $e^{u_i^*} = \varepsilon\kappa^{-1}$ and hence $i \in I^{\complement}_{\varepsilon,\kappa}$. Further, $\lambda_i^* > 0$ is equivalent to $e^{u_i^*} r_i(K)e^{v_j^*} > \kappa\mu_i$ which is satisfied when $\varepsilon^2 r_i(K) > \kappa\mu_i$. In a symmetric way we can prove the same statement for $e^{v_j^*}$. □

It is worth to note that if $e^{u_i^*} > \varepsilon\kappa^{-1}$ then $e^{u_i^*}(Ke^{v^*})_i = \kappa\mu_i$ which corresponds to one of the original Sinkhorn marginal conditions up to the scaling factor κ.

Screening with a fixed budget number of points. Recall that the approximate dual of Sinkhorn divergence is defined with respect to the parameters ε and κ. The explicit determination of its values depends on a fixed budget numbers of points, to be chosen in a priori way, in problem (5). In the sequel of the paper, we denote by $n_b \in \{1, \ldots, n\}$ and the $m_b \in \{1, \ldots, m\}$ the budget number of points to be given for resolving problem (5). Towards this end, let us define $\xi \in \mathbb{R}^n$ and $\zeta \in \mathbb{R}^m$ to be the ordered decreasing vectors of $\mu \oslash r(K)$ and $\nu \oslash c(K)$ respectively, that is $\xi_1 \geq \xi_2 \geq \cdots \geq \xi_n$ and $\zeta_1 \geq \zeta_2 \geq \cdots \geq \zeta_m$. To keep only a budget of n_b and m_b points, the parameters κ and ε satisfy $\varepsilon^2\kappa^{-1} = \xi_{n_b}$ and $\varepsilon^2\kappa = \zeta_{m_b}$. Hence

$$\varepsilon = (\xi_{n_b}\zeta_{m_b})^{1/4} \text{ and } \kappa = \sqrt{\frac{\zeta_{m_b}}{\xi_{n_b}}}, \tag{6}$$

we then obtain $|I_{\varepsilon,\kappa}| = n_b$ and $|J_{\varepsilon,\kappa}| = m_b$. Using the previous analysis, we know, in a posterior way, that any solution (u^*, v^*) of problem (5) should satisfy: $e^{u_i^*} \geq \varepsilon\kappa^{-1}$ and $e^{v_j^*} \geq \varepsilon\kappa$ for all $(i, j) \in (I_{\varepsilon,\kappa} \times J_{\varepsilon,\kappa})$, and $e^{u_i^*} = \varepsilon\kappa^{-1}$ and $e^{v_j^*} = \varepsilon\kappa$ for all $(i, j) \in (I^{\complement}_{\varepsilon,\kappa} \times J^{\complement}_{\varepsilon,\kappa})$

Basing on that facts we "screen" the feasibility domain in problem (5) to the following one:

$$\min\{\Psi_{\varepsilon,\kappa}(u, v)\} \text{ subject to } \begin{cases} e^{u_i} \geq \varepsilon\kappa^{-1}, \text{ for all } i \in I_{\varepsilon,\kappa} \text{ and } e^{u_i} = \varepsilon\kappa^{-1} \text{ for all } i \in I^{\complement}_{\varepsilon,\kappa} \\ e^{v_j} \geq \varepsilon\kappa, \text{ for all } j \in J_{\varepsilon,\kappa} \text{ and } e^{v_i} = \varepsilon\kappa \text{ for all } j \in J^{\complement}_{\varepsilon,\kappa}, \end{cases} \tag{7}$$

where

$$\Psi_{\varepsilon,\kappa}(u, v) := \sum_{i\in I_{\varepsilon,\kappa}, j\in J_{\varepsilon,\kappa}} e^{u_i}K_{ij}e^{v_j} + \varepsilon\kappa \sum_{i\in I_{\varepsilon,\kappa}, j\in J^{\complement}_{\varepsilon,\kappa}} e^{u_i}K_{ij} + \varepsilon\kappa^{-1} \sum_{i\in I^{\complement}_{\varepsilon,\kappa}, j\in J_{\varepsilon,\kappa}} K_{ij}e^{v_j}$$
$$- \kappa \sum_{i\in I_{\varepsilon,\kappa}} \mu_i u_i - \kappa^{-1} \sum_{j\in J_{\varepsilon,\kappa}} \nu_j v_j + \Xi,$$

with $\Xi = \varepsilon^2 \sum_{i\in I^{\complement}_{\varepsilon,\kappa}, j\in J^{\complement}_{\varepsilon,\kappa}} K_{ij} - \kappa\log(\varepsilon\kappa^{-1}) \sum_{i\in I^{\complement}_{\varepsilon,\kappa}} \mu_i - \kappa^{-1}\log(\varepsilon\kappa) \sum_{j\in J^{\complement}_{\varepsilon,\kappa}} \nu_j$. We refer to problem (7) as the *screened dual of Sinkhorn divergence*.

First order conditions for problem (7). Let $(u^{\text{sc}}, v^{\text{sc}})$ be an optimal solution of problem (7), then we have

$$e^{u_i^{\text{sc}}} \sum_{j\in J_{\varepsilon,\kappa}} K_{ij} e^{v_j^{\text{sc}}} + \varepsilon\kappa e^{u_i^{\text{sc}}} \sum_{j\in J^{\complement}_{\varepsilon,\kappa}} K_{ij} - \kappa\mu_i = 0, \text{ for all } i \in I_{\varepsilon,\kappa},$$

and

$$e^{v_j^{\text{sc}}} \sum_{i\in I_{\varepsilon,\kappa}} K_{ij} e^{u_i^{\text{sc}}} + \varepsilon\kappa^{-1} e^{v_j^{\text{sc}}} \sum_{i\in I^{\complement}_{\varepsilon,\kappa}} K_{ij} - \kappa^{-1}\nu_j = 0, \text{ for all } j \in J_{\varepsilon,\kappa}.$$

Therefore

$$e^{u_i^{\text{sc}}} = \frac{\kappa\mu_i}{\sum_{j\in J_{\varepsilon,\kappa}} K_{ij} e^{v_j^{\text{sc}}} + \varepsilon\kappa \sum_{j\in J^{\complement}_{\varepsilon,\kappa}} K_{ij}}, \text{ for all } i \in I_{\varepsilon,\kappa},$$

and

$$e^{v_i^{\text{sc}}} = \frac{\kappa^{-1}\nu_j}{\sum_{i\in I_{\varepsilon,\kappa}} K_{ij} e^{u_i^{\text{sc}}} + \varepsilon\kappa^{-1} \sum_{i\in I^{\complement}_{\varepsilon,\kappa}} K_{ij}}, \text{ for all } j \in J_{\varepsilon,\kappa},$$

Lemma 2. *Let (u^{sc}, v^{sc}) be an optimal solution of problem* (7)*. Then, one has*

$$\frac{\varepsilon}{\kappa} \vee \frac{\min_{i\in I_{\kappa,\varepsilon}} \mu_i}{\varepsilon|J^{\complement}_{\varepsilon,\kappa}| + \varepsilon \vee \frac{\max_{j\in J_{\kappa,\varepsilon}} \nu_j}{n\varepsilon\kappa \min_{i,j} K_{ij}} |J_{\kappa,\varepsilon}|} \le e^{u_i^{\text{sc}}} \le \frac{\varepsilon}{\kappa} \vee \frac{\max_{i\in I_{\kappa,\varepsilon}} \mu_i}{m\varepsilon \min_{i,j} K_{ij}}, \tag{8}$$

and

$$\varepsilon\kappa \vee \frac{\min_{j\in J_{\kappa,\varepsilon}} \nu_j}{\varepsilon|I^{\complement}_{\varepsilon,\kappa}| + \varepsilon \vee \frac{\kappa \max_{i\in I_{\kappa,\varepsilon}} \mu_i}{m\varepsilon \min_{i,j} K_{ij}} |I_{\kappa,\varepsilon}|} \le e^{v_j^{\text{sc}}} \le \varepsilon\kappa \vee \frac{\max_{j\in J_{\kappa,\varepsilon}} \nu_j}{n\varepsilon \min_{i,j} K_{ij}} \tag{9}$$

for all $i \in I_{\kappa,\varepsilon}$ and $j \in J_{\kappa,\varepsilon}$.

Proof. We prove only the first statement (8) and symmetrically we can prove the second statement (9) for $e^{v_j^{\text{sc}}}$. For all $i \in I_{\varepsilon,\kappa}$, we have $e^{u_i^{\text{sc}}} > \frac{\varepsilon}{\kappa}$ or $e^{u_i^{\text{sc}}} = \frac{\varepsilon}{\kappa}$. In one hand, if $e^{u_i^{\text{sc}}} > \frac{\varepsilon}{\kappa}$ then according to the optimality conditions $\lambda_i^{\text{sc}} = 0$. Then $e^{u_i^{\text{sc}}} \sum_{j=1}^m K_{ij} e^{v_j^{\text{sc}}} = \kappa\mu_i$. In another hand, we have

$$e^{u_i^{\text{sc}}} \min_{i,j} K_{ij} \sum_{j=1}^m e^{v_j^{\text{sc}}} \le e^{u_i^{\text{sc}}} \sum_{j=1}^m K_{ij} e^{v_j^{\text{sc}}} = \kappa\mu_i.$$

We further observe that $\sum_{j=1}^m e^{v_j^{\text{sc}}} = \sum_{j\in J_{\kappa,\varepsilon}} e^{v_j^{\text{sc}}} + \sum_{j\in J^{\complement}_{\varepsilon,\kappa}} e^{v_j^{\text{sc}}} \ge \varepsilon\kappa|J_{\kappa,\varepsilon}| + \varepsilon\kappa|J^{\complement}_{\varepsilon,\kappa}| = \varepsilon\kappa(|J_{\kappa,\varepsilon}| + |J^{\complement}_{\varepsilon,\kappa}|) = \varepsilon\kappa m$. Then

$$\max_{i\in I_{\kappa,\varepsilon}} e^{u_i^{\text{sc}}} \le \frac{\varepsilon}{\kappa} \vee \frac{\max_{i\in I_{\kappa,\varepsilon}} \mu_i}{m\varepsilon \min_{i,j} K_{ij}}.$$

Analogously, one can obtain for all $j \in J_{\kappa,\varepsilon}$

$$\max_{j\in J_{\kappa,\varepsilon}} e^{v_j^{\text{sc}}} \le \varepsilon\kappa \vee \frac{\max_{j\in J_{\kappa,\varepsilon}} \nu_j}{n\varepsilon \min_{i,j} K_{ij}}. \tag{10}$$

Now, since $K_{ij} \le 1$, we have

$$e^{u_i^{\text{sc}}} \sum_{j=1}^m e^{v_j^{\text{sc}}} \ge e^{u_i^{\text{sc}}} \sum_{j=1}^m K_{ij} e^{v_j^{\text{sc}}} = \kappa\mu_i.$$

Moreover, using (10), we get

$$\sum_{j=1}^m e^{v_j^{\text{sc}}} = \sum_{j\in J_{\kappa,\varepsilon}} e^{v_j^{\text{sc}}} + \sum_{j\in J^{\complement}_{\varepsilon,\kappa}} e^{v_j^{\text{sc}}} \le \varepsilon\kappa|J^{\complement}_{\varepsilon,\kappa}| + \varepsilon\kappa \vee \frac{\max_{j\in J_{\kappa,\varepsilon}} \nu_j}{n\varepsilon \min_{i,j} K_{ij}} |J_{\kappa,\varepsilon}|.$$

Therefore,

$$\min_{i\in I_{\kappa,\varepsilon}} e^{u_i^{\text{sc}}} \ge \frac{\varepsilon}{\kappa} \vee \frac{\kappa \min_{I_{\kappa,\varepsilon}} \mu_i}{\varepsilon\kappa|J^{\complement}_{\varepsilon,\kappa}| + \varepsilon\kappa \vee \frac{\max_{j\in J_{\kappa,\varepsilon}} \nu_j}{n\varepsilon \min_{i,j} K_{ij}} |J_{\kappa,\varepsilon}|}.$$

$\square$

4 Numerical experiments

References

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

[2] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc., 2013.

[3] L. Kantorovich. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 2:227–229, 1942.

[4] R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.