Sinkhorn, end to end

Duality, matrix scaling, debiasing, and convergence

This exercise derives entropic optimal transport from Young's inequality, obtains Sinkhorn scaling, proves nonnegativity of the Sinkhorn divergence, and establishes convergence modulo constants.

Part I — Setting

Let $Z \subset \mathbb{R}^d$ be compact, let α, β be probability measures on Z, and put $X = \operatorname{supp} \alpha$, $Y = \operatorname{supp} \beta$. Assume that $c : Z \times Z \to \mathbb{R}_+$ is continuous, symmetric, satisfies $c(x,x) = 0$, and is L-Lipschitz in each variable. Write $D = \operatorname{diam}(Z)$ and $\varepsilon > 0$.

Principle 1 (Entropic transport)**.**

$$\mathrm{OT}_\varepsilon(\alpha,\beta) := \min_{\pi \in \Pi(\alpha,\beta)} \left\{ \int c \,\mathrm{d}\pi + \varepsilon \mathrm{KL}(\pi \| \alpha \otimes \beta) \right\}, \qquad \mathrm{KL}(\pi\|\mu) = \int \log \frac{\mathrm{d}\pi}{\mathrm{d}\mu} \,\mathrm{d}\pi.$$

The value is $+\infty$ if $\pi \not\ll \mu$. The minimizer exists and is unique.

Principle 2 (Young's inequality)**.** For $\varphi(r) = r \log r - r + 1$,

$$rs \le \varphi(r) + e^s - 1, \qquad r \ge 0,\ s \in \mathbb{R},$$

with equality exactly when $r = e^s$.

Principle 3 (Kernel assumption for debiasing)**.** Set $k_\varepsilon(x,y) = e^{-c(x,y)/\varepsilon}$ and

$$\langle \lambda, K_\varepsilon \sigma \rangle := \iint k_\varepsilon(x,y) \,\mathrm{d}\lambda(x) \,\mathrm{d}\sigma(y), \qquad \|\lambda\|^2_{K_\varepsilon} := \langle \lambda, K_\varepsilon \lambda \rangle.$$

Part III assumes that k_ε is positive semi-definite. This holds for $c(x,y) = \|x-y\|^2$. More generally, e^{-tc} is positive semi-definite for every $t > 0$ when c is conditionally of negative type [3].

Remark (Discrete notation). For $\alpha = \sum_i a_i \delta_{x_i}$ and $\beta = \sum_j b_j \delta_{y_j}$, set $C_{ij} = c(x_i, y_j)$ and $K_{ij} = e^{-C_{ij}/\varepsilon}$. A coupling is a matrix $P \ge 0$ with $P\mathbf{1} = a$ and $P^\top \mathbf{1} = b$.

Part II — Duality and Sinkhorn scaling

Step 1 (dual problem). Let $\mu = \alpha \otimes \beta$ and $\mathrm{KL}_+(\pi\|\mu) = \int \varphi(\,\mathrm{d}\pi/\,\mathrm{d}\mu) \,\mathrm{d}\mu$.

(a) Show that $\mathrm{KL}_+ = \mathrm{KL}$ on $\Pi(\alpha,\beta)$, and justify existence and uniqueness of the primal minimizer.

(b) For $f \in C(X)$ and $g \in C(Y)$, use Young's inequality to prove

$$D(f,g) := \int f \,\mathrm{d}\alpha + \int g \,\mathrm{d}\beta - \varepsilon \iint \left(e^{(f(x)+g(y)-c(x,y))/\varepsilon} - 1 \right) \mathrm{d}\alpha(x) \,\mathrm{d}\beta(y) leq \mathrm{OT}_\varepsilon(\alpha,\beta). \quad (1)$$

(c) Show that equality holds when

$$\mathrm{d}\pi_\varepsilon = e^{(f \oplus g - c)/\varepsilon} \,\mathrm{d}(\alpha \otimes \beta) \quad (2)$$

has marginals α and β.

Step 2 (soft transforms). Define

$$\mathsf{S}_\beta(g)(x) = -\varepsilon \log \int_Y e^{(g(y)-c(x,y))/\varepsilon}\, \mathrm{d}\beta(y), \quad \mathsf{S}_\alpha(f)(y) = -\varepsilon \log \int_X e^{(f(x)-c(x,y))/\varepsilon}\, \mathrm{d}\alpha(x). \tag{3}$$

(a) Show that $f = \mathsf{S}_\beta(g)$ maximizes $D(\cdot, g)$ and that $g = \mathsf{S}_\alpha(f)$ maximizes $D(f, \cdot)$.

(b) Show that these equations are the two marginal constraints for (2).

(c) Check $D(f + k, g - k) = D(f, g)$ and $\mathsf{S}_\beta(g + k) = \mathsf{S}_\beta(g) - k$.

Step 3 (matrix scaling). Sinkhorn iteration is $f_{k+1} = \mathsf{S}_\beta(g_k)$, $g_{k+1} = \mathsf{S}_\alpha(f_{k+1})$. In the discrete setting, let $u_i = e^{f_i/\varepsilon}$ and $v_j = e^{g_j/\varepsilon}$.

(a) Derive

$$u \leftarrow \frac{1}{K(v \odot b)}, \qquad v \leftarrow \frac{1}{K^\top (u \odot a)}, \qquad P_{ij} = a_i b_j K_{ij} u_i v_j.$$

(b) Show that each half-step enforces one marginal and increases the dual value.

(c) Characterize a fixed point.

Part III — Sinkhorn divergence

Define

$$\mathsf{S}_\varepsilon(\alpha, \beta) := \mathrm{OT}_\varepsilon(\alpha, \beta) - \frac{1}{2}\mathrm{OT}_\varepsilon(\alpha, \alpha) - \frac{1}{2}\mathrm{OT}_\varepsilon(\beta, \beta).$$

Step 4 (entropic bias). Take $Z = \{0, 1\}$, $c(x, y) = |x - y|^2$, and $\alpha = \frac{1}{2}(\delta_0 + \delta_1)$.

(a) Find a constant Schrödinger pair and prove $\mathrm{OT}_\varepsilon(\alpha, \alpha) = \varepsilon \log \frac{2}{1+e^{-1/\varepsilon}} > 0$.

(b) Compute the limits as $\varepsilon \to 0$ and $\varepsilon \to \infty$.

Step 5 (nonnegativity after debiasing). Assume that k_ε is positive semi-definite. Let f_α be a symmetric optimal potential for $\mathrm{OT}_\varepsilon(\alpha, \alpha)$ and set $\lambda_\alpha = e^{f_\alpha/\varepsilon}\alpha$; define f_β, λ_β similarly.

(a) Show

$$K_\varepsilon \lambda_\alpha = e^{-f_\alpha/\varepsilon} \quad \alpha\text{-a.e.}, \qquad \|\lambda_\alpha\|_{K_\varepsilon}^2 = 1, \qquad \frac{1}{2}\mathrm{OT}_\varepsilon(\alpha, \alpha) = \int f_\alpha \,\mathrm{d}\alpha.$$

(b) Insert (f_α, f_β) into the cross dual problem and prove

$$\mathsf{S}_\varepsilon(\alpha, \beta) \geq \frac{\varepsilon}{2}\|\lambda_\alpha - \lambda_\beta\|_{K_\varepsilon}^2 \geq 0. \tag{4}$$

(c) Why does positive semi-definiteness not necessarily imply separation?

Part IV — Convergence

For $h \in C(X)$ set $\|h\|_{\circ,\infty} = \frac{1}{2}\operatorname{osc}(h) = \frac{1}{2}(\sup h - \inf h)$, and similarly on $C(Y)$.

Step 6 (quotient by constants). Prove $\inf_{a\in\mathbb{R}} \|h-a\|_\infty = \frac{1}{2}\operatorname{osc}(h)$. Deduce that $\|\cdot\|_{\circ,\infty}$ is the norm of the Banach quotient $C(X)/\mathbb{R}$. Why is this quotient necessary?

Step 7 (contraction of one soft transform). For $g_t = g_0 + t(g_1 - g_0)$, prove

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{S}_\beta(g_t)(x) = -\int_Y (g_1 - g_0)\,\mathrm{d}\nu_{t,x}, \quad \nu_{t,x}(\,\mathrm{d}y) = \frac{e^{(g_t(y)-c(x,y))/\varepsilon}\beta(\,\mathrm{d}y)}{\int_Y e^{(g_t - c(x,\cdot))/\varepsilon}\,\mathrm{d}\beta}. \tag{5}$$

Using that $\nu_{t,x} - \nu_{t,x'}$ has mass zero, show

$$\|\mathsf{S}_\beta(g_1) - \mathsf{S}_\beta(g_0)\|_{\circ,\infty} \le \kappa_\beta \|g_1 - g_0\|_{\circ,\infty}, \qquad \kappa_\beta := \frac{1}{2}\sup_{x,x',t}\|\nu_{t,x} - \nu_{t,x'}\|_{\mathrm{TV}}. \tag{6}$$

Step 8 (linear convergence). For $\nu_w = e^w\beta/\int e^w\,\mathrm{d}\beta$, prove

$$\|\nu_u - \nu_v\|_{\mathrm{TV}} \le 2\big(1 - e^{-\operatorname{osc}(u-v)}\big). \tag{7}$$

Deduce $\kappa_\alpha, \kappa_\beta \le 1 - e^{-2LD/\varepsilon} < 1$. Show that $T = \mathsf{S}_\beta \circ \mathsf{S}_\alpha$ contracts $C(X)/\mathbb{R}$, and conclude that Sinkhorn converges linearly to a unique Schrödinger pair modulo $(f,g) \mapsto (f+k, g-k)$.

Solutions

Solution. **Step 1.** For $r = \mathrm{d}\pi/\,\mathrm{d}\mu$,

$$\mathrm{KL}_+(\pi\|\mu) = \mathrm{KL}(\pi\|\mu) - \pi(Z^2) + \mu(Z^2) = \mathrm{KL}(\pi\|\mu).$$

The coupling set is weakly compact; the objective is lower semicontinuous and finite at $\alpha\otimes\beta$. Strict convexity of entropy gives uniqueness.

Apply Young's inequality with $s = (f \oplus g - c)/\varepsilon$ and integrate against μ. The marginal constraints give $\int (f\oplus g)\,\mathrm{d}\pi = \int f\,\mathrm{d}\alpha + \int g\,\mathrm{d}\beta$, hence (1). Equality holds exactly for the Gibbs density (2). If it has the prescribed marginals, it is primal and dual optimal.

Solution. **Step 2.** For $h \in C(X)$,

$$\partial_f D(f,g)[h] = \int_X h(x)\left[1 - int_Y e^{(f(x)+g(y)-c(x,y))/\varepsilon}\,\mathrm{d}\beta(y)\right]\,\mathrm{d}\alpha(x).$$

It vanishes for every h exactly when $f = \mathsf{S}_\beta(g)$ on X; concavity gives the maximum. The symmetric calculation gives $g = \mathsf{S}_\alpha(f)$. These two equations are precisely the marginal constraints for the Gibbs plan. Finally, direct substitution gives both constant-shift identities.

Solution. **Step 3.** Equation (3) gives $u_i^{-1} = \sum_j K_{ij}v_jb_j$ and $v_j^{-1} = \sum_i K_{ij}u_ia_i$. The Gibbs plan is $P_{ij} = a_ib_jK_{ij}u_iv_j$. The u-update gives $P\mathbf{1} = a$; the v-update gives $P^\top\mathbf{1} = b$. Each update maximizes one dual variable. At a fixed point both marginals hold, so the potentials are dual optimal and P is the primal optimizer.

Solution. **Step 4.** By symmetry take $f = g = \phi$ constant. The marginal equation is

$$e^{2\phi/\varepsilon} \frac{1 + e^{-1/\varepsilon}}{2} = 1.$$

Since $\mathrm{OT}_\varepsilon(\alpha, \alpha) = 2\phi$, $\mathrm{OT}_\varepsilon(\alpha, \alpha) = \varepsilon \log \frac{2}{1+e^{-1/\varepsilon}} > 0$. It tends to 0 as $\varepsilon \to 0$ and to $\iint c \,\mathrm{d}\alpha \,\mathrm{d}\alpha = 1/2$ as $\varepsilon \to \infty$.

Solution. **Step 5.** The self-marginal equation gives $K_\varepsilon \lambda_\alpha = e^{-f_\alpha/\varepsilon}$ α-a.e. Therefore

$$\|\lambda_\alpha\|_{K_\varepsilon}^2 = 1, \qquad \mathrm{OT}_\varepsilon(\alpha, \alpha) = 2 \int f_\alpha \,\mathrm{d}\alpha.$$

Evaluating the cross dual at (f_α, f_β) gives

$$\mathrm{OT}_\varepsilon(\alpha, \beta) \geq \int f_\alpha \,\mathrm{d}\alpha + int f_\beta \,\mathrm{d}\beta - \varepsilon(\langle \lambda_\alpha, K_\varepsilon \lambda_\beta \rangle - 1).$$

Subtracting the self-costs yields

$$\mathsf{S}_\varepsilon(\alpha, \beta) \geq \varepsilon(1 - \langle \lambda_\alpha, K_\varepsilon \lambda_\beta \rangle) = \frac{\varepsilon}{2} \|\lambda_\alpha - \lambda_\beta\|_{K_\varepsilon}^2 \geq 0.$$

A positive semi-definite kernel defines only a seminorm. Separation requires the kernel to be characteristic; Gaussian kernels satisfy this stronger condition [1].

Solution. **Step 6.** $\|h - a\|_\infty = \max(\sup h - a, a - \inf h)$ is minimized at $a = (\sup h + \inf h)/2$, with value $\mathrm{osc}(h)/2$. Thus $\|\cdot\|_{\circ,\infty}$ is the quotient norm on $C(X)/\mathbb{R}$, a Banach space. The quotient removes the undetermined constant in the dual potentials.

Solution. **Step 7.** Differentiating (3) gives (5). Integrating in t and subtracting the values at x, x' gives

$$[\mathsf{S}_\beta(g_1) - \mathsf{S}_\beta(g_0)](x) - [\mathsf{S}_\beta(g_1) - \mathsf{S}_\beta(g_0)](x') = -\int_0^1 \langle g_1 - g_0, \nu_{t,x} - \nu_{t,x'} \rangle \,\mathrm{d}t.$$

The signed measure has mass zero. Subtract any constant from $g_1 - g_0$, minimize its sup norm, and take one half of the supremum over x, x' to obtain (6).

Solution. **Step 8.** Shift $u - v$ so that $0 \leq u - v \leq M := \mathrm{osc}(u - v)$. For $h = \mathrm{d}\nu_u / \mathrm{d}\nu_v$,

$$h = \frac{e^{u-v}}{\int e^{u-v} \,\mathrm{d}\nu_v} \geq e^{-M}.$$

Since both measures have mass one,

$$\frac{1}{2} \|\nu_u - \nu_v\|_{\mathrm{TV}} = \int (1 - h)_+ \,\mathrm{d}\nu_v \leq 1 - e^{-M},$$

proving (7). For $\nu_{t,x}, \nu_{t,x'}$, the Gibbs potentials differ by $[c(x', \cdot) - c(x, \cdot)]/\varepsilon$, whose oscillation is at most $2LD/\varepsilon$. Hence $\kappa_\alpha, \kappa_\beta < 1$ with the stated bounds.

Thus $T = \mathsf{S}_\beta \circ \mathsf{S}_\alpha$ contracts $C(X)/\mathbb{R}$ with factor $q = \kappa_\alpha \kappa_\beta$. Banach's theorem gives a unique fixed class. If $T(f_*) = f_* + k$ and $g_* = \mathsf{S}_\alpha(f_*)$, the Gibbs plan has total mass 1 from its second marginal and $e^{-k/\varepsilon}$ from its first; hence $k = 0$. Therefore (f_*, g_*) is the unique Schrödinger pair modulo constants and

$$\|T^n f - T^n f_*\|_{\circ,\infty} \leq q^n \|f - f_*\|_{\circ,\infty}.$$

The elementary estimate gives $n = \mathcal{O}(e^{2LD/\varepsilon} \log(1/\delta))$ iterations for accuracy δ as $\varepsilon \to 0$; sharper Hilbert-metric constants are available [2].

Remark. Young's inequality gives the dual and Gibbs plan; coordinate maximization gives Sinkhorn scaling; quotient-space contraction gives existence and convergence. Debiasing is a separate argument based on positive semi-definiteness of $e^{-c/\varepsilon}$.

References

[1] J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouvé and G. Peyré, *Interpolating between optimal transport and MMD using Sinkhorn divergences*, Proceedings of AISTATS, PMLR, 2019, pp. 2681–2690.

[2] J. Franklin and J. Lorenz, *On the scaling of multidimensional matrices*, Linear Algebra and its Applications **114–115** (1989), 717–735.

[3] I. J. Schoenberg, *Metric spaces and positive definite functions*, Transactions of the American Mathematical Society **44** (1938), no. 3, 522–536.