Optimal Transport

John Thickstun

Let $(\mathcal{X}, p)$ and $(\mathcal{Y}, q)$ be finite probability spaces with $|\mathcal{X}| = n$ and $|\mathcal{Y}| = m$. Let $\Pi(p,q) \subset \Delta^{m\times n}$ be the collection of distributions on the product space $\mathcal{X}\times\mathcal{Y}$ with marginals p on $\mathcal{X}$ and q on $\mathcal{Y}$. Consider a cost $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}_+$ and the optimal transport problem

$$d_c(p,q) = \min_{\pi\in\Pi(p,q)} \langle c, \pi\rangle = \min_{\pi\in\Pi(p,q)} \sum_{x,y} c(x,y)\pi(x,y). \tag{1}$$

For example, if $\mathcal{X},\mathcal{Y} \subset \mathbb{R}^n$ and $c(x,y) = \|x-y\|_2$ is the standard Euclidean distance, then the optimal transport between p and q is the (1-)Wasserstein distance

$$d_c(p,q) = \min_{\pi\in\Pi(p,q)} \sum_{x,y} \|x-y\|_2 \pi(x,y) = W_1(p,q). \tag{2}$$

Concretely, if $\mathbf{1}_n \in \mathbb{R}^n, \mathbf{1}_m \in \mathbb{R}^m$ are the vectors of all-ones, then the constraints on the marginals can be written as

$$p(x) = \sum_{i=1}^m \pi(x, y_i) = (\pi\mathbf{1}_m)_x, \text{ and } q(y) = \sum_{i=1}^n \pi(x_i, y) = (\mathbf{1}_n^T\pi)_y. \tag{3}$$

And we can write the optimal transport problem as

$$d_c(p,q) = \min_{\substack{\pi\mathbf{1}_m = p\\ \pi^\top\mathbf{1}_n = q}} \sum_{x,y} c(x,y)\pi(x,y). \tag{4}$$

In the context of e.g. the Wasserstein GAN, it can be helpful to think of the discrete Wasserstein distance (and more generally, the optimal transport) between two finite distributions p and q as being a minibatch approximation of the Wasserstein distance between continuous distributions. If p,q are continuous distributions on $\mathbb{R}^d$, $x_1,\dots,x_n \sim p$, and $y_1,\dots,y_m \sim q$, denote the empirical distributions over samples by $\tilde{p}$ and $\tilde{q}$ respectively:

$$\tilde{p}(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{x_i = x}, \quad \tilde{q}(y) = \frac{1}{m}\sum_{i=1}^m \mathbf{1}_{y_i = y}, \tag{5}$$

We can approximate $W_1(p,q) \approx W_1(\tilde{p},\tilde{q})$.

Entropy-Regularized Optimal Transport

The optimization problem given by Equation (4) is a linear program. We could solve this problem using an LP-solver, but we can make our life a lot easier if allow for some approximation error.

Instead of solving (4), we'll instead solve an entropy-regularized optimal transport problem, where $H(\pi)$ is the entropy of the joint distribution [Cuturi, 2013]:

$$d_c^\lambda(p,q) = \min_{\pi \in \Pi(p,q)} \langle c, \pi \rangle - \lambda H(\pi). \tag{6}$$

As $\lambda \to 0$, $d_c^\lambda(p,q) \to d_c(p,q)$ and the optimal solution to $d_c^\lambda(p,q)$ converges to the minimizer of $d_c(p,q)$ with highest entropy.

Analogous to how we can reframe maximum likelihood estimation as KL-divergence minimization, we can also reframe (6) as a KL-divergence minimization problem. Define $k(x,y) \equiv e^{-c(x,y)/\lambda}$. If $Z_\lambda = \sum_{x,y} k(x,y)$ then $\frac{1}{Z_\lambda} k(x,y)$ defines a (Gibbs) probability distribution p_k^λ and

$$D(\pi \parallel p_k^\lambda) = \sum_{x,y} \pi(x,y) \log \frac{\pi(x,y) Z_\lambda}{k(x,y)} = \frac{1}{\lambda} \langle c, \pi \rangle - H(\pi) + \log Z_\lambda.$$

And it follows that

$$\arg\min_{\pi \in \Pi(p,q)} \langle c, \pi \rangle - \lambda H(\pi) = \arg\min_{\pi \in \Pi(p,q)} D(\pi \parallel p_k^\lambda). \tag{7}$$

By compactness of $\Pi(p,q)$ and strong convexity of the negative entropy, d_c^λ has a unique minimizer π_λ, which can be interpreted geometrically as the information projection of the cost matrix's associated Gibbs distribution at temperature λ onto $\Pi(p,q)$.

A Primal Algorithm

We can recover π_λ using iterative I-projections. Let $\Pi(p)$ and $\Pi(q)$ denote the row and column marginal constraints, so in particular $\Pi(p,q) = \Pi(p) \cap \Pi(q)$. Initialize $\pi_\lambda^{(0)} = p_k^\lambda$ and define the alternating projections

$$\pi_\lambda^{(\ell+1)} \equiv \begin{cases} \arg\min_{\pi \in \Pi(p)} D(\pi \parallel \pi_\lambda^{(\ell)}) & \ell \text{ even}, \\ \arg\min_{\pi \in \Pi(q)} D(\pi \parallel \pi_\lambda^{(\ell+1)}) & \ell \text{ odd}. \end{cases}$$

Because $\Pi(p)$ and $\Pi(q)$ are affine sets, $\pi_\lambda^{(\ell)} \to \pi_\lambda$ by classical convex analysis.

Without loss of generality, suppose ℓ is even. Then $\pi_\lambda^{(\ell+1)}$ satisfies

$$\frac{\partial}{\partial \pi} \left[D(\pi \parallel \pi_\lambda^{(\ell)}) - \langle f, \pi \mathbf{1}_m - p \rangle \right] = 0 \text{ (first-order optimality)}.$$

And for a particular pair (x,y),

$$1 + \log \pi_{\lambda,f}^{(\ell+1)}(x,y) - \log \pi_\lambda^{(\ell)}(x,y) - f_x = 0.$$

Therefore $\pi_{\lambda,f}^{(\ell+1)}(x,y) = e^{f_x - 1} \pi_\lambda^{(\ell)}(x,y)$ and because $\pi_\lambda^{(\ell+1)} \in \Pi(p)$ we must have

$$e^{f_x - 1} = \frac{p(x)}{\sum_y \pi_\lambda^{(\ell)}(x,y)}.$$

Packaging up this and the analogous reasoning for odd ℓ, we have

$$\pi_\lambda^{(2\ell)} = \operatorname{diag}\left(\frac{p}{\pi_\lambda^{(2\ell-1)} \mathbf{1}_m} \right) \pi_\lambda^{(2\ell-1)}, \text{ and } \pi_\lambda^{(2\ell+1)} = \operatorname{diag}\left(\frac{q}{\mathbf{1}_n^\top \pi_\lambda^{(2\ell)}} \right) \pi_\lambda^{(2\ell)}.$$

Sinkhorn's Algorithm

The preceding algorithm iterated on primal variables $\pi_\lambda^{(\ell)}$. It turns out we can iterate more efficiently on dual variables, by exploiting the following structure of the optimal solution.

Proposition. *Let $K \in \mathbb{R}^{n\times m}$ with $K_{x,y} = k(x,y)$. For some $u \in \mathbb{R}^n$, $v \in \mathbb{R}^m$,*

$$\pi_\lambda = \operatorname{diag}(u) K \operatorname{diag}(v). \tag{8}$$

Proof. Introduce dual variables $f \in \mathbb{R}^n$ and $g \in \mathbb{R}^m$ and consider the Lagrangian

$$\mathcal{L}(\pi, f, g) = \langle c, \pi\rangle - \lambda H(\pi) - \langle f, \pi \mathbf{1}_m - p\rangle - \langle g, \pi^\top \mathbf{1}_n - q\rangle. \tag{9}$$

First order optimality occurs for π_λ satisfying

$$c(x,y) + \lambda \log \pi_\lambda(x,y) - f_x - g_y = 0. \tag{10}$$

In other words,

$$\pi_\lambda(x,y) = e^{f_x/\lambda - 1/2} e^{-c(x,y)/\lambda} e^{g_y/\lambda - 1/2}. \tag{11}$$

□

The constraints $\Pi(p,q)$ determine the values u and v. In particular, the row and column sums of π_λ must match those of $p \otimes q$. Finding u and v is known as the matrix scaling problem, in the sense that we want to scale K's rows and columns to match the row and column sums of $p \otimes q$. Unpacking the problem a bit, we want u,v that satisfy

$$p = \pi_\lambda \mathbf{1}_m = \operatorname{diag}(u)(Kv) \text{ and } q = \pi_\lambda^\top \mathbf{1}_n = \operatorname{diag}(v)(K^\top u). \tag{12}$$

Sinkhorn's algorithm approximates a solution to these equations by initializing $u^{(1)} \equiv \mathbf{1}_n$, $v^{(1)} \equiv \mathbf{1}_m$, and constructing the sequence

$$u^{(\ell+1)} \equiv \frac{p}{Kv^{(\ell)}}, \text{ and } v^{(\ell+1)} \equiv \frac{q}{K^\top u^{(\ell+1)}}. \tag{13}$$

Division here is interpreted entry-wise.

This is equivalent to our previous primal algorithm. Consider primal iterates

$$\begin{aligned}\tilde{\pi}_\lambda^{(2\ell)} &\equiv \operatorname{diag}(u^{(\ell+1)}) K \operatorname{diag}(v^{(\ell)}),\\ \tilde{\pi}_\lambda^{(2\ell+1)} &\equiv \operatorname{diag}(u^{(\ell+1)}) K \operatorname{diag}(v^{(\ell+1)}).\end{aligned}$$

Rearranging terms, observe that

$$K \operatorname{diag}(v^{(\ell)}) = \frac{\tilde{\pi}_\lambda^{(2\ell-1)}}{\operatorname{diag}(u^{(\ell)})}.$$

It follows that

$$\begin{aligned}\tilde{\pi}_\lambda^{(2\ell)} = \operatorname{diag}(u^{(\ell+1)}) K \operatorname{diag}(v^{(\ell)}) = \operatorname{diag}\left(\frac{p}{Kv^{(\ell)}}\right) \frac{\tilde{\pi}_\lambda^{(2\ell-1)}}{\operatorname{diag}(u^{(\ell)})}\\ = \operatorname{diag}\left(\frac{p}{\operatorname{diag}(u^{(\ell)})Kv^{(\ell)}}\right) \tilde{\pi}_\lambda^{(2\ell-1)} = \operatorname{diag}\left(\frac{p}{\tilde{\pi}_\lambda^{(2\ell-1)}\mathbf{1}_m}\right) \tilde{\pi}_\lambda^{(2\ell-1)}.\end{aligned}$$

Likewise,

$$\mathrm{diag}(u^{(\ell+1)})K = \frac{\tilde{\pi}_\lambda^{(2\ell)}}{\mathrm{diag}(v^{(\ell)})}.$$

And similarly we see that

$$\tilde{\pi}_\lambda^{(2\ell+1)} \equiv \mathrm{diag}(u^{(\ell+1)})K\,\mathrm{diag}(v^{(\ell+1)}) = \frac{\tilde{\pi}_\lambda^{(2\ell)}}{\mathrm{diag}(v^{(\ell)})}\,\mathrm{diag}\left(\frac{q}{K^\top u^{(\ell+1)}}\right)$$

$$= \mathrm{diag}\left(\frac{q}{\mathrm{diag}(v^{(\ell)})K^\top u^{(\ell+1)}}\right)\tilde{\pi}_\lambda^{(2\ell)} = \mathrm{diag}\left(\frac{q}{\mathbf{1}_n^\top \tilde{\pi}_\lambda^{(2\ell)}}\right)\tilde{\pi}_\lambda^{(2\ell)}.$$

Therefore the Sinkhorn dual algorithm is identical to the primal algorithm.

References

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, 2013. (document)