1 Optimal Transport

The following summarizes some chapters of computational optimal transport book [1].

1.1 Monge's and Kantorovich's formulation

Given $a, b \in \triangle^{n-1}$, where $\triangle^{n-1}$ is unit simplex. $\alpha = \sum_{i=1}^{n} a_i \delta_{x_i}, \beta = \sum_{j=1}^{m} b_j \delta_{y_i}$ are discrete measures. The optimal transport problem tries to find a map that associate each point x_i to a single point y_j such that masses are preserved. This corresponds to classical Monge's formulation of optimal transport

$$\min_{T:\mathcal{X}\to\mathcal{Y}:T_{\#}\alpha=\beta} \sum_{i=1}^{n} c(x_i, T(x_i)) \tag{1}$$

where $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is cost defined over support of measures, and that the constraints is simply that T is constrained to be a pushforward from α to β. The problem with Monge's formulation is that the problem is combinatorial and nonconvex, so hard to solve. Kantorovich relax the deterministic nature of transport map, allowing mass at each source point split and be dispatched to multiple target points. This information is encoded in $P \in \mathbb{R}_{+}^{n\times m}$, where P_{ij} describes amount of mass flowing from x_i to y_j. The Kantorovich formulation of optimal transport for discrete measures is then

$$\min_{P\in\mathbb{R}_{+}^{n\times m}:P1_m=a,\, P^T1_n=b} \langle C, P\rangle \tag{2}$$

where the constraints specifies the set of admissible transport map to be a coupling of marginals α, β and $C \in \mathbb{R}^{n\times m}$ is the cost matrix, i.e. $C_{ij} = c(x_i, y_j)$. The optimization problem is a linear program and hence can be easily solved using simplex algorithm. In addition, we can instead solve the dual problem, and because of zero duality gap, equivalently solves the primal problem,

$$\max_{f,g\in\mathbb{R}^n\times\mathbb{R}^m:f_i+g_j\le C_{ij}} \langle f, a\rangle + \langle g, b\rangle \tag{3}$$

where f, g are called dual potential.

1.2 Entropic Regularization

Regularizing the original optimal transport problem brings computational and statistical benefits. In particular, the optimization problem can now be solved with fast matrix scaling algorithms that scales with strength of regularization. In addition, the sample efficiency for regularized problem is also superior. The entropy of coupling between two 1 dimensional discrete measure is given by

$$H(P) = -\sum_{i,j} P_{ij}(\log(P_{ij}) - 1) = -\langle P, \log P\rangle + 1^T P 1 \tag{4}$$

where taking logarithm and subtraction are elementwise operations. The entropic regularized problem

$$\min_{P\in\mathbb{R}_{+}^{n\times m}:P1_m=a,P^T1_n=b} \langle P, C\rangle + \epsilon H(P) \tag{5}$$

We can interpret primal objective as the information projection of the Gibbs kernel $K \in \mathbb{R}^{n\times m}$ where $K_{ij} = e^{\frac{C_{ij}}{\epsilon}}$ onto the admissible couplings $U(a, b) = \left\{P \in \mathbb{R}_{+}^{n\times m} \mid P1_m = a,\, P^T1_n = b\right\}$.

$$\langle P, C\rangle - \epsilon \langle P, \log P)\rangle + \epsilon 1^T P 1 = \epsilon \left\langle P, \log\left(P \oslash e^{-C/\epsilon}\right)\right\rangle - \epsilon 1^T P 1 + \epsilon 1^T K 1 = \epsilon \mathrm{KL}(P\|K) \tag{6}$$

We can solve regularized optimal transport problem with Sinkhorn algorithm [2]. The basic idea is to write the 1st order optimality condition for the primal variables for the Lagrangian,

$$\mathcal{L}(P, f, g) = \langle P, C \rangle - \epsilon H(P) - \langle f, P1_m - a \rangle - \left\langle g, P^T 1_n - b \right\rangle \tag{7}$$

$$\partial \mathcal{L}(P, f, g)/\partial P_{ij} = C_{ij} + \epsilon \log(P_{ij}) - f_i - g_j = 0 \quad \Rightarrow \quad P_{ij} = e^{(-C_{ij}+f_i+g_j)/\epsilon} \tag{8}$$

Note optimal P can be written as scaling of Gibbs kernel row-wise by $u = e^{f/\epsilon}$ and column-wise by $v = e^{g/\epsilon}$, i.e. $P = \operatorname{diag}(u) K \operatorname{diag}(v)$. Substitute expression for P into marginal constraints, get $a = \operatorname{diag}(u) K \operatorname{diag}(v) 1_m = u \odot (Kv)$ and $b = v \odot (K^T u)$. The Sinkhorn's algorithm updates u, v alternatingly

$$u^{(\ell+1)} \leftarrow a \oslash (Kv^{(\ell)}) \qquad v^{(\ell+1)} \leftarrow b \oslash (K^T u^{(\ell+1)}) \tag{9}$$

Each iteration of the algorithm uses $\mathcal{O}(nm)$ computation for matrix-vector products and can be accelerated to $\mathcal{O}(n \log n)$ using convolution if suppport is over gridded space. See Figure (1) to visualize effect of varying the regularization strength.

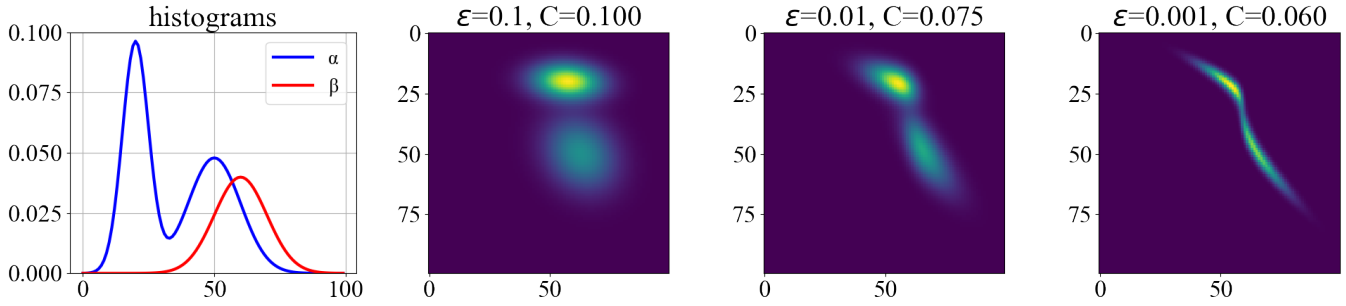

Figure 1: Entropic regularized optimal transport wrt 1d histograms

Sinkhorn's iteration can be numerically unstable as $\epsilon \to 0$ due zeros in K, resulting in null values during division by zero. One solution is to perform computation of u, v in the log-domain [3, 4]. This is equivalent to block coordinate ascent on the dual of entropic regularized optimal transport problem (5),

$$\max_{f,g \in \mathbb{R}^{n \times m}} \langle f, a \rangle + \langle g, b \rangle - \epsilon \left\langle e^{f/\epsilon}, K e^{g/\epsilon} \right\rangle \tag{10}$$

Then $0 = \partial L_{\text{dual}}(f, g)/\partial f = a - e^{f/\epsilon} \odot (Ke^{g/\epsilon})$ implies $f = \epsilon \log(a) - \epsilon \log(Ke^{g/\epsilon})$ and similarly for g. So,

$$f^{(\ell+1)} \leftarrow \epsilon \log(a) - \epsilon \log(Ke^{g^{(\ell)}/\epsilon}) \qquad g^{(\ell+1)} \leftarrow \epsilon \log(b) - \epsilon \log(Ke^{f^{(\ell+1)}/\epsilon}) \tag{11}$$

This is equivalent to Sinkhorn's iteration in log domain. Additionally, we can define a numerically stable softmin operator based on log-sum-exp, $\operatorname{softmin}_\epsilon(z) = -\epsilon \log \sum_i e^{-z_i/\epsilon} = -\epsilon \mathrm{LSE}(-z/\epsilon)$. Note $f_i = \epsilon \log a_i - \epsilon \log \sum_j e^{-(C_{ij}-g_j)/\epsilon} = \operatorname{softmin}_\epsilon(C_{ij} - f_i - g_j) + f_i + \epsilon \log a_i$. Therefore, (11) can be equivalently written as,

$$f^{(\ell+1)} \leftarrow \operatorname{softmin}_\epsilon(S(f^{(\ell)}, g^{(\ell)})) + f^{(\ell)} + \epsilon \log(a) \tag{12}$$

$$g^{(\ell+1)} \leftarrow \operatorname{softmin}_\epsilon(S(f^{(\ell+1)}, g^{(\ell)})) + g^{(\ell)} + \epsilon \log(b) \tag{13}$$

where $S(f, g)_{ij} = C_{ij} - f_i - g_j$. Figure (1) is computed using log domain stabilization, where choice of $\epsilon = .001$ would yield null values without log domain stabilization.

1.3 Unbalanced Transport

We can extend optimal transport to general measures by enforcing soft constraints on the coupling [3],

$$\min_{P \in \mathbb{R}_+^{n \times m}} \langle P, C \rangle + \epsilon H(P) + \rho \mathrm{KL}(P 1_m \| a) + \rho \mathrm{KL}(P^T 1_m \| b) \tag{14}$$

Similar to balanced case, we can show that optimal coupling is again a scaling of K, i.e. $P = \mathrm{diag}(u) K \, \mathrm{diag}(v)$ where $u = (a \oslash (Kv))^\lambda$ and $v = (b \oslash (K^T u))^\lambda$ where $\lambda = \frac{\rho}{\rho + \epsilon}$ [5]. We can derive a similar Sinkhorn iteration in the log domain for unbalanced case.

References

[1] Gabriel Peyré and Marco Cuturi. "Computational Optimal Transport". In: *arXiv:1803.00567 [stat]* (Mar. 18, 2020). arXiv: 1803.00567. URL: http://arxiv.org/abs/1803.00567 (visited on 03/05/2021).

[2] Marco Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances". In: *arXiv:1306.0895 [stat]* (June 4, 2013). arXiv: 1306.0895. URL: http://arxiv.org/abs/1306.0895 (visited on 10/14/2020).

[3] Lenaic Chizat et al. "Scaling Algorithms for Unbalanced Transport Problems". In: *arXiv:1607.05816 [math]* (May 22, 2017). arXiv: 1607.05816. URL: http://arxiv.org/abs/1607.05816 (visited on 04/27/2021).

[4] Bernhard Schmitzer. "Stabilized Sparse Scaling Algorithms for Entropy Regularized Transport Problems". In: *arXiv:1610.06519 [cs, math]* (Feb. 11, 2019). arXiv: 1610.06519. URL: http://arxiv.org/abs/1610.06519 (visited on 04/27/2021).

[5] Charlie Frogner et al. "Learning with a Wasserstein Loss". In: *arXiv:1506.05439 [cs, stat]* (Dec. 29, 2015). arXiv: 1506.05439. URL: http://arxiv.org/abs/1506.05439 (visited on 10/16/2020).