

# NOTES ON ENTROPIC REGULARIZATION

JIAFENG CHEN AND FRANCISCO RIVERA

## 1. STATEMENT

The original problem is

$$\begin{aligned} \min \quad & \sum_{i,j} P_{ij} C_{ij} \\ \text{s.t.} \quad & \mathbf{P} \mathbf{1} = \mathbf{a} \\ & \mathbf{P}^T \mathbf{1} = \mathbf{b} \end{aligned}$$

Denote the feasible region by  $\mathbf{U}(\mathbf{a}, \mathbf{b})$ . *Entropic regularization* considers the following modified problem:

$$L_{\mathbf{C}}^{\epsilon}(\mathbf{a}, \mathbf{b}) = \min_{\mathbf{P} \in \mathbf{U}(\mathbf{a}, \mathbf{b})} \sum_{i,j} P_{ij} C_{ij} + \underbrace{\epsilon \sum_{i,j} P_{ij} (\log P_{ij} - 1)}_{-H(\mathbf{P})}.$$

We can show that the negative entropy  $-H(\mathbf{P})$  is 1-strongly convex (i.e. Hessian minus  $I$  is positive semi-definite). Thus the objective is  $\epsilon$ -strongly convex. The strong convexity means that the optimum  $\mathbf{P}_{\epsilon}$  is unique. Unsurprisingly,  $\mathbf{P}_{\epsilon} \rightarrow \mathbf{P}^{\star}$  as  $\epsilon \rightarrow 0$  where  $\mathbf{P}^{\star}$  is the maximal-entropy solution to the original problem  $L_{\mathbf{C}}^0$ . Moreover, note that if  $\epsilon \rightarrow \infty$ , we are essentially maximizing the entropy, and unsurprisingly, the maximal entropy solution is  $\mathbf{a}\mathbf{b}^T$ : In terms of probability, maximal entropy joint distribution is assuming independence.

We can reformulate the problem via *KL projection*. Note that minimizing the regularized objective is akin to minimizing

$$\begin{aligned} & \sum_{i,j} \frac{1}{\epsilon} P_{ij} C_{ij} + P_{ij} (\log P_{ij} - 1) + e^{-C_{ij}/\epsilon} \\ &= \sum_{i,j} P_{ij} \left( \log P_{ij} - \log \left( e^{-C_{ij}/\epsilon} \right) \right) - P_{ij} + e^{-C_{ij}/\epsilon} \\ &= \sum_{i,j} P_{ij} \log \frac{P_{ij}}{K_{ij}} - P_{ij} + K_{ij} =: \text{KL} \left( \frac{\mathbf{P}}{\mathbf{K}} \right), \end{aligned}$$

where  $\mathbf{K}_{ij} = \exp(-C_{ij}/\epsilon)$  is called a *Gibbs kernel*. Thus the problem is akin to projecting  $\mathbf{K}$  onto  $\mathbf{U}(\mathbf{a}, \mathbf{b})$  via the KL-divergence as a distance metric.

The Lagrangian of the regularized problem is

$$\begin{aligned} \text{Lagrange}(\mathbf{P}, \mathbf{f}, \mathbf{g}) &= \sum_{i,j} P_{ij} C_{ij} + \epsilon \sum_{i,j} P_{ij} (\log P_{ij} - 1) \\ &\quad + \sum_i f_i \left( a_i - \sum_j P_{ij} \right) + \sum_j g_j \left( b_j - \sum_i P_{ij} \right), \end{aligned}$$

for which the first-order conditions yield

$$C_{ij} + \epsilon \log P_{ij} - f_i - g_j = 0.$$

Rewriting yields

$$P_{ij} = e^{f_i/\epsilon} e^{-C_{ij}/\epsilon} e^{g_j/\epsilon} = u_i K_{ij} v_j$$

for some  $\mathbf{u}, \mathbf{v}$ . The condition prescribed by  $\mathbf{U}(\mathbf{a}, \mathbf{b})$ , written via  $\mathbf{u}, \mathbf{v}$ , yield

$$\mathbf{u} \odot (\mathbf{K}\mathbf{v}) = \mathbf{a} \quad \mathbf{v} \odot (\mathbf{K}^T \mathbf{u}) = \mathbf{b}.$$

Since the optimum is unique, we need only find  $\mathbf{u}, \mathbf{v}$  such that these conditions hold.

Given  $\mathbf{v}$ , we can compute  $\mathbf{u}$  via  $\mathbf{u} = \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}}$  and vice versa. This motivates Sinkhorn's algorithm by initializing  $\mathbf{v}^{(0)} = \mathbf{1}$  and computing

$$\mathbf{u}^{(\ell+1)} = \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}^{(\ell)}} \quad \mathbf{v}^{(\ell+1)} = \frac{\mathbf{b}}{\mathbf{K}^T \mathbf{u}^{(\ell+1)}}$$

It has been shown (2017!) that Sinkhorn updates achieves a  $\tau$ -approximate solution of the unregularized problem in  $O(n^2 \log(n) \tau^{-3})$  update iterations.