§3 Discrete OT: Entropic Regularization*

Jiayao Zhang†

November 1, 2019

In the last meeting, we proved (two thirds of) the **fundamental theorem of optimal transport**, where we note that being optimal is indeed an instrinsic property that depends on the support of the transportation plan. With this theorem we were able to recover a special case of the **Brenier's theorem**: in $\mathbb{R}^d$ when Monge map exists, it must be the gradient of some convex function. Let us first recall the theorem.

Theorem 1. ***Fundamental theorem of optimal transport.***

If c is continuous and bounded from below and for some $f \in L_1(\mu)$, $g \in L_1(\nu)$ we have for all $(x, y) \in X \times Y$,

$$c(x, y) \leq f(x) + g(y), \tag{1}$$

then TFAE:

(a) $\gamma \in \Gamma(\mu, \nu)$ *is optimal.*

(b) $\operatorname{supp}(\gamma)$ *is c-cyclically monotone.*

(c) There exists a c-concave function φ *such that* $\varphi^+ \in L_1(\mu)$ *and* $\operatorname{supp}(\gamma) \subset \partial^{c+}\varphi$.

In this meeting, we will move on to discuss several methods for solving optimal transport when the underlying space is discrete and of finite cardinality.

1 Dual ascent method

Recall the dual formulation of Kantorovich's problem,

$$\begin{aligned} \max \quad & \int \varphi \, \mathrm{d}\mu + \int \psi \, \mathrm{d}\nu \\ \text{subject to} \quad & \varphi(x) + \psi(y) \leq c(x, y), \quad \forall (x, y) \in X \times Y, \\ & \varphi \in L^1(\mu), \quad \psi \in L^1(\nu). \end{aligned} \tag{2}$$

In the discrete case, φ, μ, μ, ν are all vectors, c is encoded in the cost matrix $\boldsymbol{C}$ where $C_{ij} = c(i, j)$, and the coupling $\gamma \in \Pi(\mu, \nu)$ is represented by the coupling matrix $\boldsymbol{P}$ where the sum over the i-th row gives $\mu(i)$; and j-th column $\nu(j)$. For simplicity, we use the notations interchangably.

*This note is based on [PC+19] and [AG13], and was presented at Penn optimal transport reading group.

†University of Pennsylvania, `jiayaozhang@acm.org`.

The primal problem, minimum weight bipartite matching, can be solved by Hungarian algorithm, and the dual problem can be cast into a maximum flow problem that we are familiar with: we can solve integer flow exactly using Ford-Fulkerson, or say Edmonds-Karp, in time polynomial in the number of dual variables.

2 Some notions from information theory

We will focus in this section discrete probability spaces, the case for continuous probability spaces are sometimes *not* analogous and often induces headaches. In this section, we will write random variables in capital letters, and use caligraphical letters or Ω for the underlying sample spaces.

Let X be an r.v. on $(\Omega, \mathcal{F}, \mathbb{P})$ with $|\Omega| = n < \infty$ and write $p_i = \mathbb{P}(X = i)$. The **entropy** of X is defined as

$$H(X) = -\sum_{i \in \mathcal{X}} p_i \log p_i \geq 0, \tag{3}$$

which quantifies the uncertainty in the random variable. Subject to moment constraints of X (e.g., $\mathbb{E}X^k = \gamma_k$), the entropy-maximizing distribution will always be exponential family with the moment as its sufficient statistics. For example, the uniform distribution on Ω has the largest entropy of $\log|\Omega|$; in all distributions with prescribed second moment, Guassians with scale σ has the largest *differential entropy* $\frac{1}{2}\ln(2\pi\sigma)$, which unlike discrete case, can be in general negative.

Given two r.v.s X on $\mathcal{X}$ and Y on $\mathcal{Y}$, we can discuss their **joint entropy** $H(X, Y)$, **conditional entropy** $H(X|Y)$, and if they are defined on the same probability space Ω, **cross entropy** $H(X; Y)$, **relative entropy** (or KL divergence) $D(X\|Y)$, and **mutual information** $I(X; Y)$. Denote by $p(\cdot)$ the pmf, we have

$$\begin{aligned}
H(X, Y) &= -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log p(x, y) \\
H(X|Y) &= \sum_{y \in \mathcal{Y}} p(y) H(X|Y = y) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log p(x|y) = H(X, Y) - H(Y) \\
H(X; Y) &= -\sum_{x \in \Omega} \mathbb{P}(X = x) \log \mathbb{P}(Y = x) \\
D(X\|Y) &= \sum_{x \in \Omega} \mathbb{P}(X = x) \log \frac{\mathbb{P}(X = x)}{\mathbb{P}(Y = x)} = H(X; Y) - H(X) \\
I(X; Y) &= D(p(X, Y)\|p(X)p(Y)) = H(X) + H(Y) - H(X, Y) = H(X) - H(X|Y).
\end{aligned} \tag{4}$$

We now remark a few properties.

- By Jensen's inequality, $D(X\|Y) \geq 0$ and is zero iff X and Y are equal in distribution. It follows that $I(X; Y) \geq 0$ and is zero iff X and Y are independent.
- Noting the Hessian of $H(X)$, $\nabla^2 H(X)$ is negative (semi)definite, hence $H(\cdot)$ is concave. Furthermore, noting

  $$\nabla^2 H(X) = -\operatorname{diag}(p_i) \quad \Rightarrow \quad \nabla^2 H(X) - 1 \cdot \boldsymbol{I} \prec 0, \tag{5}$$

  since $p_i = \mathbb{P}(X = i) < 1$. That is, $H(\cdot)$ is 1-concave.

3 Entropic regularization for Kantorovich's problem

Given a coupling matrix $\boldsymbol{P}$, we can define

$$H(\boldsymbol{P}) = -\sum_{ij} P_{ij} \log P_{ij}, \tag{6}$$

and consider the regularized Kantorovich's problem:

$$\min_{\gamma \in \Pi(\mu,\nu)} \langle \gamma, c \rangle - \epsilon H(\gamma) = \mathrm{tr}(\boldsymbol{P}^\top \boldsymbol{C}) - \epsilon H(\boldsymbol{P}). \tag{7}$$

Noting the objective is ϵ-convex, it has a unique optimal solution $\boldsymbol{P}_\epsilon$. Furthermore, we have the following theorem regarding its behaviour as we vary ϵ.

Theorem 2. ***Convergence with*** ϵ**.** *As* $\epsilon \to 0$*,*

$$\boldsymbol{P}_\epsilon \to \arg\min_{\boldsymbol{P} \in \Pi(\mu,\nu)} \{-H(\boldsymbol{P}) : \mathrm{tr}(\boldsymbol{P}^\top \boldsymbol{C}) = OPT\}. \tag{8}$$

As $\epsilon \to \infty$*,*

$$\boldsymbol{P}_\epsilon \to \mu \otimes \nu. \tag{9}$$

Proof. Note since μ and ν are fixed, regularizing on $H(\boldsymbol{P})$ is equivalent to regularizing on the mutual information between $\pi^X_\# \gamma$ and $\pi^Y_\# \gamma$, hence as $\epsilon \to 0$, $\boldsymbol{P}_\epsilon$ is pushed to be $\mu \otimes \nu$. On the other hand, as $\epsilon \to 0$, noting for any optimal plan $\boldsymbol{P}$,

$$\mathrm{tr}(\boldsymbol{P}^\top \boldsymbol{C}) \leq \mathrm{tr}(\boldsymbol{P}_\epsilon^\top \boldsymbol{C}), \quad \mathrm{tr}(\boldsymbol{P}_\epsilon^\top \boldsymbol{C}) - \mathrm{tr}(\boldsymbol{P}^\top \boldsymbol{C}) \leq \epsilon \left(H(\boldsymbol{P}_\epsilon) - H(\boldsymbol{P}) \right), \tag{10}$$

which implies

$$0 \leq \mathrm{tr}(\boldsymbol{P}_\epsilon^\top \boldsymbol{C}) - \mathrm{tr}(\boldsymbol{P}^\top \boldsymbol{C}) \leq \epsilon \left(H(\boldsymbol{P}_\epsilon) - H(\boldsymbol{P}) \right). \tag{11}$$

Hence as $\epsilon \to 0$, $\boldsymbol{P}_\epsilon \to \boldsymbol{P}^*$ (limit by the compactness of the set of coupling wrt the narrow topology) such that $\boldsymbol{P}^*$ is optimal and is entropy-maximizing among all optimal plans. □

References

[AG13] Luigi Ambrosio and Nicola Gigli. A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, pages 1–155. Springer, 2013.

[PC+19] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.