Self-study: Semi-discrete Optimal Transport

Tianpei Xie

Aug. 19th., 2022

Contents

1 Optimal transport, entropy regularization, c-transform

- For discrete measures, $\alpha = \sum_{i=1}^{n} a_i \delta_{\boldsymbol{x}_i}$ and $\beta := \sum_{i=1}^{m} b_i \delta_{\boldsymbol{y}_i}$, the primal problem for optimal transport is

$$\min_{\boldsymbol{P} \in \mathbb{R}_{+}^{n\times m}} \langle \boldsymbol{P}\,,\, \boldsymbol{C} \rangle = \sum_{i,j} C_{i,j} P_{i,j} \tag{1}$$
$$\text{s.t. } \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a} \tag{2}$$
$$\boldsymbol{P}^T \mathbf{1}_n = \boldsymbol{b} \tag{3}$$
$$P_{i,j} \geq 0$$

  where $\boldsymbol{C}_{n,m} := [C_{i,j}]_{i\in[1:n], j\in[1:m]}$, $C_{i,j} := c(\boldsymbol{x}_i, \boldsymbol{y}_j) \geq 0$. The feasible set is defined as

$$U(\boldsymbol{a}, \boldsymbol{b}) := \left\{ \boldsymbol{P} \in \mathbb{R}_{+}^{n\times m} : \boldsymbol{P}\mathbf{1}_m = \boldsymbol{a},\ \boldsymbol{P}^T \mathbf{1}_n = \boldsymbol{b} \right\} \tag{4}$$

- the corresponding dual problem with respect to primal problem is

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}^n, \boldsymbol{\mu} \in \mathbb{R}^m} \langle \boldsymbol{\lambda}\,,\, \boldsymbol{a} \rangle + \langle \boldsymbol{\mu}\,,\, \boldsymbol{b} \rangle \tag{5}$$
$$\text{s.t. } \lambda_i + \mu_j \leq C_{i,j} \quad \forall\, i \in [1:n], j \in [1:m] \tag{6}$$

  where $\boldsymbol{\lambda} = [\lambda_i]_n$, $\boldsymbol{\mu} = [\mu_j]_m$ are **dual variables** (slack variables) for marginal distribution constrain $\boldsymbol{a}$ and $\boldsymbol{b}$. We denote $\boldsymbol{\lambda} \oplus \boldsymbol{\mu} := \boldsymbol{\lambda}\mathbf{1}_m^T + \mathbf{1}_n \boldsymbol{\mu}^T \in \mathbb{R}^{n\times m}$ so that the linear constraints is $\boldsymbol{\lambda} \oplus \boldsymbol{\mu} \leq \boldsymbol{C}$. Such dual variables $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ are often referred to as "***Kantorovich potentials***." The feasible set of the dual problem is defined as

$$R(\boldsymbol{C}) := \{ \boldsymbol{\lambda} \in \mathbb{R}^n, \boldsymbol{\mu} \in \mathbb{R}^m : \boldsymbol{\lambda} \oplus \boldsymbol{\mu} \leq \boldsymbol{C} \} \tag{7}$$

  where $\boldsymbol{\lambda} \oplus \boldsymbol{\mu} = \boldsymbol{\lambda}\mathbf{1}_m + \mathbf{1}_n \boldsymbol{\mu}^T$.

- The **probability interpretation** of orignal primal and dual Kantorovich optimal transport problem:

$$(P) \quad \mathcal{L}_c(\alpha, \beta) = \min_{(X,Y)\sim\pi} \mathbb{E}_{(X,Y)}\left[c(X,Y) \right] \tag{8}$$
$$\text{s.t. } X \sim \alpha,$$
$$Y \sim \beta$$

$$(D) \quad \mathcal{L}_c(\alpha, \beta) = \max_{(\lambda,\mu) \in \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y})} \mathbb{E}_{X\sim\alpha}\left[\lambda(X) \right] + \mathbb{E}_{Y\sim\beta}\left[\mu(Y) \right] \tag{9}$$
$$\text{s.t. } \lambda(x) + \mu(y) \leq c(x,y), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y},$$

- We also have the **entropic regularized optimal transport problem**

$$L_{\boldsymbol{C}}^{\epsilon}(\boldsymbol{a}, \boldsymbol{b}) = \min_{\boldsymbol{P} \in U(\boldsymbol{a}, \boldsymbol{b})} \langle \boldsymbol{P}\,,\, \boldsymbol{C} \rangle - \epsilon H(\boldsymbol{P}) \tag{10}$$

  where the second term is entropy

$$H(\boldsymbol{P}) := -\sum_{i,j} P_{i,j} \left(\log(P_{i,j}) - 1 \right) \tag{11}$$

This problem has a unique optimal solution (maximum entropy optimal transport plan)

$$\boldsymbol{P}^* = \text{diag}\left(\boldsymbol{u}\right) \boldsymbol{K} \text{diag}\left(\boldsymbol{v}\right) \tag{12}$$

where $\boldsymbol{u} = [\exp(\lambda_i/\epsilon)] = \exp(\boldsymbol{\lambda}/\epsilon)$ and $\boldsymbol{v} = [\exp(\mu_j/\epsilon)] = \exp(\boldsymbol{\mu}/\epsilon)$.

- The dual problem of the **maximum entropy optimal transport problem**:

$$L^{\epsilon}_{\boldsymbol{C}}(\boldsymbol{a}, \boldsymbol{b}) = \max_{\boldsymbol{\lambda}\in\mathbb{R}^n, \boldsymbol{\mu}\in\mathbb{R}^m} \langle \boldsymbol{\lambda}\,,\, \boldsymbol{a}\rangle + \langle \boldsymbol{\mu}\,,\, \boldsymbol{b}\rangle - \epsilon \left\langle \exp\left(\boldsymbol{\lambda}/\epsilon\right)\,,\, \boldsymbol{K} \exp\left(\boldsymbol{\mu}/\epsilon\right)\right\rangle \tag{13}$$

  where $\boldsymbol{K} = \exp\left(-\boldsymbol{C}/\epsilon\right)$ is the Gibbs distribution.

- The **probability interpretation** of primal and dual maximum entropy optimal transport problem:

$$\begin{aligned} (P) \quad \mathcal{L}^{\epsilon}(\alpha, \beta) := \min_{(X,Y)\sim\pi} & \; \mathbb{E}_{(X,Y)}\left[c(X, Y)\right] + \epsilon\, I(X; Y) \\ \text{s.t. } & \; X \sim \alpha \\ & \; Y \sim \beta \end{aligned} \tag{14}$$

  where $I(X; Y) := \mathbb{KL}\left(\pi \,\|\, \alpha \otimes \beta\right)$ is the mutual information between X and Y.

$$\begin{aligned} (D) \quad \mathcal{L}^{\epsilon}(\alpha, \beta) := \sup_{(\lambda,\mu)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} & \; \mathbb{E}_{X\sim\alpha}\left[\lambda(X)\right] + \mathbb{E}_{Y\sim\beta}\left[\mu(Y)\right] \\ & - \epsilon\mathbb{E}_{X\sim\alpha, Y\sim\beta}\left[\exp\left(\frac{-c(X, Y) + \lambda(X) + \mu(Y)}{\epsilon}\right)\right]. \end{aligned} \tag{15}$$

- Given a cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, $f : \mathcal{X} \to \mathbb{R}$, the c***-transform*** of f is defined as

$$f^c(y) := \inf_{x\in\mathcal{X}} c(x, y) - f(x) \tag{16}$$

  The function $f^c : \mathcal{Y} \to \mathbb{R}$ is also called the c***-conjugate function*** of f. For discrete case, we have C***-transform vector*** for cost matrix $\boldsymbol{C} = [C_{i,j}]_{n\times m}$ and vector $\boldsymbol{f} = [f_1, \ldots, f_n] \in \mathbb{R}^n$,

$$\boldsymbol{f}^{\boldsymbol{C}}_j := \min_{i\in[1:n]} C_{i,j} - \boldsymbol{f}_i \tag{17}$$

  The vector $\boldsymbol{f}^{\boldsymbol{C}} \in \mathbb{R}^m$ is also called the C***-conjugate vector*** of $\boldsymbol{f}$.

  Similarly, $g : \mathcal{Y} \to \mathbb{R}$, the $\bar{c}$***-transform*** of g is defined as

$$g^{\bar{c}}(x) := \inf_{y\in\mathcal{Y}} c(x, y) - g(y) \tag{18}$$

  For discrete case, we have $\bar{C}$***-transform vector*** $\boldsymbol{g}^{\bar{\boldsymbol{C}}} \in \mathbb{R}^n$ for cost matrix $\boldsymbol{C} = [C_{i,j}]_{n\times m}$ and vector $\boldsymbol{g} = [g_1, \ldots, g_m] \in \mathbb{R}^m$,

$$\boldsymbol{g}^{\bar{\boldsymbol{C}}}_i := \min_{j\in[1:m]} C_{i,j} - \boldsymbol{g}_j \tag{19}$$

  A function $\psi : \mathcal{X} \to \mathbb{R}$ is c**-concave** if there exists some function $\phi : \mathcal{Y} \to \mathbb{R}$ and cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ so that ψ is the $\bar{c}$-transform of ϕ, i.e. $\psi = \phi^{\bar{c}}$. Denote ψ as c-concave($\mathcal{X}$).

  A function $\phi : \mathcal{Y} \to \mathbb{R}$ is $\bar{c}$**-concave** if there exists some function $\psi : \mathcal{X} \to \mathbb{R}$ and cost function $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ so that ϕ is the c-transform of ψ, i.e. $\phi = \psi^c$. Denote ϕ as $\bar{c}$-concave($\mathcal{Y}$)

  For distance $c = d$, $f^c = f^{\bar{c}}$, thus we drop their distinctions.

- Suppose that $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ is real valued.
  1. For any $f_1 : \mathcal{X} \to \mathbb{R}$ and $f_2 : \mathcal{X} \to \mathbb{R}$, $f_1 \leq f_2, \Leftrightarrow \; f_1^c \geq f_2^c$
  2. For any $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{Y} \to \mathbb{R}$, $f^{c\bar{c}} \geq f$, $g^{\bar{c}c} \geq g$ In general, $f^{c\bar{c}}$ is the **smallest** c**-concave function** larger than f
  3. $f^{c\bar{c}c} = f^c$ and $g^{\bar{c}c\bar{c}} = g^{\bar{c}}$; in other words, $f^{c\bar{c}} = f$ if and only if f is a c-concave function
- **Proposition 1.1** *If $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a distance, then the function $f : \mathcal{X} \to \mathbb{R}$ is* c***-concave*** *if and only if f is* ***Lipschitz continuous*** *with Lipschitz constant less than 1 w.r.t. the distance c. We will denote by Lip_1 the set of these functions. Moreover, for every $f \in Lip_1$, i.e. $\|f\|_L \leq 1$, we have the c-transform of f, $f^c = -f$. [Santambrogio, 2015]*
- Thus the dual problem (5) is equivalent to an **unconstrained optimization problem**

$$\mathcal{L}_c(\alpha, \beta) := \max_{\lambda \in \mathcal{C}(\mathcal{X})} \int_{\mathcal{X}} \lambda d\alpha + \int_{\mathcal{Y}} \lambda^c d\beta \tag{20}$$

$$= \max_{\mu \in \mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} \mu^{\bar{c}} d\alpha + \int_{\mathcal{Y}} \mu d\beta \tag{21}$$

2 Semidiscrete optimal transport

2.1 semidiscrete problem formulation

Given discrete measure $\beta := \sum_{i=1}^m b_i \delta_{\boldsymbol{y}_i}$, the $\bar{c}$-transform of dual potential $\boldsymbol{\mu}$ is defined by restricting the minimization to the support $(\boldsymbol{y}_i)$ of β

$$\boldsymbol{\mu}^{\bar{c}}(\boldsymbol{x}) = \min_{j=1,\ldots,m} \left(c(\boldsymbol{x}, \boldsymbol{y}_j) - \mu_j\right), \quad \forall \boldsymbol{x} \in \mathcal{X}, \forall \boldsymbol{\mu} \in \mathbb{R}^m \tag{22}$$

Note that this is imposing that the support of β is equal to $\mathcal{X}$. The $\bar{c}$-transform map a vector $\boldsymbol{\mu}$ to $\boldsymbol{\mu}^{\bar{c}}(\boldsymbol{x}) \in \mathcal{C}(\mathcal{X})$, a smooth function on $\mathcal{X}$.

With $\bar{c}$-transform, we can apply the unconstrained dual formulation (21) and the problem becomes

$$\begin{aligned}\mathcal{L}_c(\alpha, \beta) &:= \max_{\mu \in \mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} \mu^{\bar{c}} d\alpha + \int_{\mathcal{Y}} \mu d\beta \\ &= \max_{\boldsymbol{\mu} \in \mathbb{R}^m} \int_{\mathcal{X}} \boldsymbol{\mu}^{\bar{c}}(\boldsymbol{x}) d\alpha(\boldsymbol{x}) + \langle \boldsymbol{\mu} \, , \, \boldsymbol{b} \rangle \end{aligned} \tag{23}$$

We can define the **Laguerre cells** associated to the dual weights $\boldsymbol{\mu}$

$$\begin{aligned}\mathbb{L}_j(\boldsymbol{\mu}) &:= \left\{\boldsymbol{x} \in \mathcal{X} : c(\boldsymbol{x}, \boldsymbol{y}_j) - \mu_j \leq c(\boldsymbol{x}, \boldsymbol{y}_{j'}) - \mu_{j'}, \;\; \forall j' \neq j\right\} \\ &= \left\{\boldsymbol{x} \in \mathcal{X} : \boldsymbol{\mu}^{\bar{c}}(\boldsymbol{x}) = c(\boldsymbol{x}, \boldsymbol{y}_j) - \mu_j\right\}\end{aligned} \tag{24}$$

We see that $\mathcal{X} = \bigcup_{j=1}^m \mathbb{L}_j$, also $\mathbb{L}_j \cap \mathbb{L}_{j'} = \emptyset$. Therefore $\{\mathbb{L}_j, j = 1, \ldots, m\}$ is a **partition** of $\mathcal{X}$. When $\boldsymbol{\mu}$ is constant, the Laguerre cells decomposition corresponds to the ***Voronoi diagram*** *partition* of the space. Each cell corresponds to a discrete mass $(b_j, \boldsymbol{y}_j)$ of β.

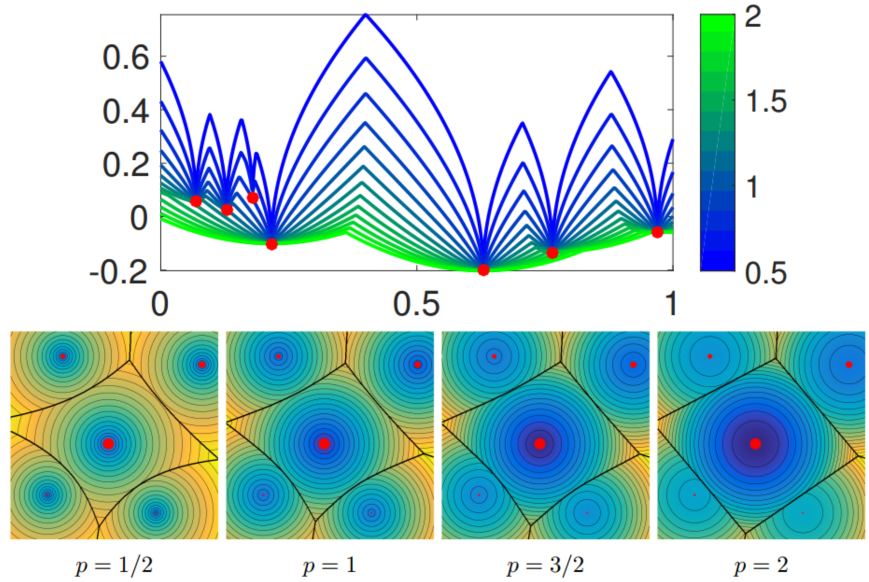

Figure 5.1: Top: examples of semidiscrete $\bar{c}$-transforms $\mathbf{g}^{\bar{c}}$ in one dimension, for ground cost $c(x, y) = |x - y|^p$ for varying p (see colorbar). The red points are at locations $(y_j, -\mathbf{g}_j)_j$. Bottom: examples of semidiscrete $\bar{c}$-transforms $\mathbf{g}^{\bar{c}}$ in two dimensions, for ground cost $c(x, y) = \|x - y\|_2^p = (\sum_{i=1}^d |x_i - y_i|)^{p/2}$ for varying p. The red points are at locations $y_j \in \mathbb{R}^2$, and their size is proportional to $\mathbf{g}_j$. The regions delimited by bold black curves are the Laguerre cells $(\mathbb{L}_j(\mathbf{g}))_j$ associated to these points $(y_j)_j$.

(a)

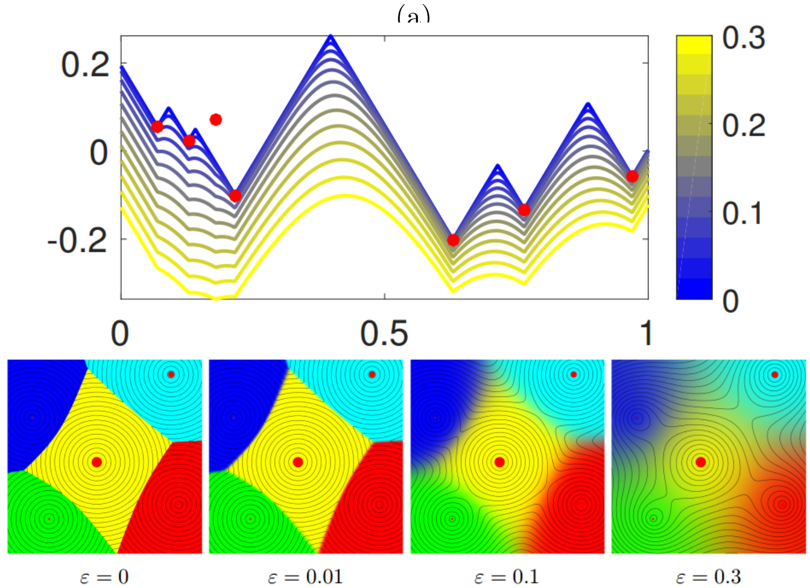

Figure 5.3: Top: examples of entropic semidiscrete $\bar{c}$-transforms $\mathbf{g}^{\bar{c},\varepsilon}$ in one dimension, for ground cost $c(x, y) = |x - y|$ for varying ε (see colorbar). The red points are at locations $(y_j, -\mathbf{g}_j)_j$. Bottom: examples of entropic semidiscrete $\bar{c}$-transforms $\mathbf{g}^{\bar{c},\varepsilon}$ in two dimensions, for ground cost $c(x, y) = \|x - y\|_2$ for varying ε. The black curves are the level sets of the function $\mathbf{g}^{\bar{c},\varepsilon}$, while the colors indicate the smoothed indicator function of the Laguerre cells χ_j^ε. The red points are at locations $y_j \in \mathbb{R}^2$, and their size is proportional to $\mathbf{g}_j$.

(b)

Figure 1: (a) The $\bar{c}$-transform of semidiscrete measures. (b)The $\bar{c}$-transform of semidiscrete measures with entropic regularization

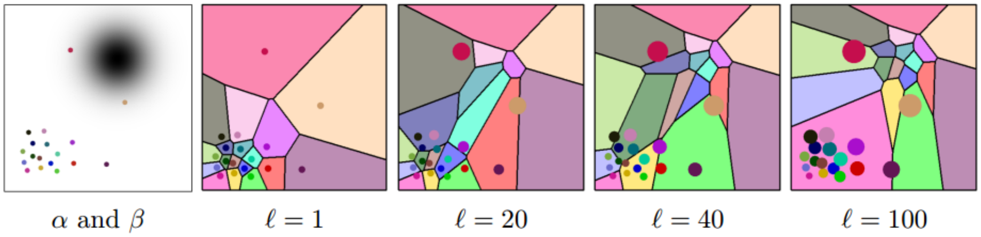

Figure 5.2: Iterations of the semidiscrete OT algorithm minimizing (5.8) (here a simple gradient descent is used). The support $(y_j)_j$ of the discrete measure β is indicated by the colored points, while the continuous measure α is the uniform measure on a square. The colored cells display the Laguerre partition $(\mathbb{L}_j(\mathbf{g}^{(\ell)}))_j$ where $\mathbf{g}^{(\ell)}$ is the discrete dual potential computed at iteration ℓ.

Figure 2: The iteration of gradient descent algorithm and the change of Laguerre cells [Peyr and Cuturi, 2019]

This allows one to conveniently rewrite the minimized energy in (23) as

$$\mathcal{E}(\boldsymbol{\mu}) := \sum_{j=1}^{m} \int_{\mathbb{L}_j(\boldsymbol{\mu})} \left(c(\boldsymbol{x}, \boldsymbol{y}_j) - \mu_j\right) d\alpha(\boldsymbol{x}) + \langle \boldsymbol{\mu} \,, \, \boldsymbol{b} \rangle \tag{25}$$

The gradient of this objective function is

$$\nabla_{\boldsymbol{\mu}} \mathcal{E}(\boldsymbol{\mu}) = \left[- \int_{\mathbb{L}_j(\boldsymbol{\mu})} d\alpha(\boldsymbol{x}) + b_j \right]_{j=1}^{m} \tag{26}$$

We can see that the gradient of objective w.r.t. dual potential $\boldsymbol{\mu}$ is the **difference** between the **discrete measure** b_j at location $\boldsymbol{y}_j$ and the probabilty **measure of Laguerre** $\mathbb{L}_j(\boldsymbol{\mu})$ associated with $(b_j, \boldsymbol{y}_j)$ (i.e. **hard assignment**). Figure 1 shows the Laguerre cell partition of the space. Given this simple form of gradient, we can directly compute the solution using gradient descent. Figure 2 shows the iterations of gradient desent algorithm and its change of Laguerre cells.

In the special case $c(x, y) = \|x - y\|_2^2$, the decomposition in Laguerre cells is also known as a ***power diagram***, which is a concept in ***computational geometry***. The cells are polyhedral and can be computed efficiently using computational geometry algorithms; see [Aurenhammer, 1987]. The most widely used algorithm relies on the fact that the power diagram of points in $\mathbb{R}^d$ is equal to the projection on $\mathbb{R}^d$ of the convex hull of the set of points $(\boldsymbol{y}_j, \|\boldsymbol{y}_j\|_2^2 - g_j)_{j=1}^m \subset R^{d+1}$. The semidiscrete OT solver can be used in computational geometry. It is also used for solving the Monge-Ampère equation.

2.2 K-means via semi-discrete optimal transport

The k-means algorithm can be re-formulated using the semi-discrete optimal transport. In particular, $\beta = \sum_{i=1}^{k} a_i \delta_{\boldsymbol{c}_i}$ is constrained to be a **discrete measure** with a finite support of **size up to** k. β is continous on domain $\mathcal{X} = \mathbb{R}^d$, $c(x, y) = \|x - y\|_2^2$, we can find β via solving the minimum Kantorvich distance estimation

$$\min_{\beta \in \mathcal{M}_{k,1}(\mathcal{X})} \mathcal{L}_c(\beta, \alpha) = \mathcal{L}_c(\alpha, \beta)$$

Indeed, one can easily show that the ***centroids*** output $\{\boldsymbol{c}_i, i = 1, \ldots, k\}$ by the k-means problem correspond to the **support** of the solution α and that its ***weights*** a_i correspond to the **fraction** of points in β *assigned to each centroid* [Canas and Rosasco, 2012].

One can show that approximating $\mathcal{L} \approx \mathcal{L}^\epsilon$ using entropic regularization results in smoothed out assignments that appear in **soft-clustering** variants of k-means, such as *mixtures of Gaussians.*

2.3 Entropic regularization

Recall from (15) that the dual of entropic regularized optimal transport

$$\mathcal{L}^\epsilon(\alpha, \beta) := \max_{(\lambda,\mu)\in\mathcal{C}(\mathcal{X})\times\mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} \lambda d\alpha + \int_{\mathcal{Y}} \mu d\beta - \epsilon \int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{-c + \lambda \oplus \mu}{\epsilon}\right) d\alpha d\beta \tag{27}$$

Similarly to the unregularized problem (9), one can minimize explicitly with respect to either λ or μ in (27), which yields a ***smoothed*** ***c*-transform**

$$\lambda^{c,\epsilon}(y) = -\epsilon \log \int_{\mathcal{X}} \exp\left(\frac{-c(x, y) + \lambda(x)}{\epsilon}\right) d\alpha(x), \quad \forall y \in \mathcal{Y} \tag{28}$$

$$\mu^{\bar{c},\epsilon}(y) = -\epsilon \log \int_{\mathcal{Y}} \exp\left(\frac{-c(x, y) + \mu(y)}{\epsilon}\right) d\beta(y), \quad \forall x \in \mathcal{X} \tag{29}$$

Compare (16) and (18) with (28) and (1), we see that instead of using min operation, in smooth c-transform, we use the <u>soft-min$^\epsilon$</u> operator soft-min$^\epsilon(\boldsymbol{z}; \boldsymbol{b}) = -\epsilon \log \sum_i b_i \exp(-z_i/\epsilon)$ to maintain smoothness of the function.

In the case of a discrete measure $\beta := \sum_{i=1}^m b_i \delta_{\boldsymbol{y}_i}$, the problem simplifies as with (22) to a finite-dimensional problem expressed as a function of the discrete dual potential $\boldsymbol{\mu}$:

$$\begin{aligned}\boldsymbol{\mu}^{\bar{c},\epsilon}(\boldsymbol{x}) &= \text{soft-min}^\epsilon_{j=1,\ldots,m} \left(c(\boldsymbol{x}, \boldsymbol{y}_j) - \mu_j; \boldsymbol{b}\right), \quad \forall \boldsymbol{x} \in \mathcal{X}, \forall \boldsymbol{\mu} \in \mathbb{R}^m \\ &= -\epsilon \log \sum_{j=1}^m b_j \exp\left(\frac{-(c(\boldsymbol{x}, \boldsymbol{y}_j) - \mu_j)}{\epsilon}\right)\end{aligned} \tag{30}$$

Similar to (23), we can solve the uncontrained dual problem using the smooth $\bar{c}$-transfrom

$$\mathcal{L}^\epsilon_c(\alpha, \beta) = \max_{\boldsymbol{\mu}\in\mathbb{R}^m} \int_{\mathcal{X}} \boldsymbol{\mu}^{\bar{c},\epsilon}(\boldsymbol{x}) d\alpha(\boldsymbol{x}) + \langle \boldsymbol{\mu}\,,\, \boldsymbol{b} \rangle \tag{31}$$

The minimized energy is

$$\begin{aligned}\mathcal{E}^\epsilon(\boldsymbol{\mu}) &:= \left\{\int_{\mathcal{X}} \boldsymbol{\mu}^{\bar{c},\epsilon}(\boldsymbol{x}) d\alpha(\boldsymbol{x}) + \langle \boldsymbol{\mu}\,,\, \boldsymbol{b} \rangle\right\} \\ &= -\mathbb{E}_\alpha \left[\epsilon \log \sum_{j=1}^m b_j \exp\left(\frac{-(c(X, \boldsymbol{y}_j) - \mu_j)}{\epsilon}\right)\right] + \langle \boldsymbol{\mu}\,,\, \boldsymbol{b} \rangle\end{aligned} \tag{32}$$

This is actually the *expectation* of negative loss function of **multiclass logistic regression problem** w.r.t. α. Note that the LR need to minimize the loss and this is to maximize the energy. Therefore, the dual representation of ***semidiscrete optimal transport*** with **entropic regularization** is ***equivalent*** to **multi-class logistic regression**.

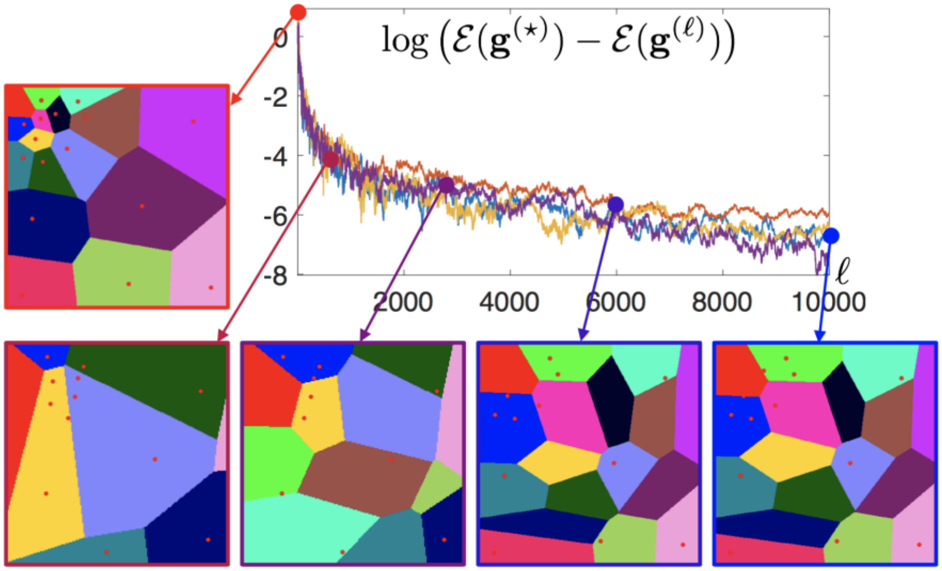

Figure 5.4: Evolution of the energy $\mathcal{E}^{\varepsilon}(\mathbf{g}^{(\ell)})$, for $\varepsilon = 0$ (no regularization) during the SGD iterations (5.14). Each colored curve shows a different randomized run. The images display the evolution of the Laguerre cells $(\mathbb{L}_j(\mathbf{g}^{(\ell)}))_j$ through the iterations.

Figure 3: The iteration of stochastic gradient descent algorithm and the change of Laguerre cells [Peyr and Cuturi, 2019]

The gradient of this functional reads

$$\nabla_{\boldsymbol{\mu}} \mathcal{E}^{\epsilon}(\boldsymbol{\mu}) = \left[- \int_{\mathcal{X}} \sigma_j^{\epsilon}(\boldsymbol{x}) d\alpha(\boldsymbol{x}) + b_j \right]_{j=1}^{m} \tag{33}$$

$$\text{where } \sigma_j^{\epsilon}(\boldsymbol{x}) = \frac{\exp\left(\frac{-(c(\boldsymbol{x}, \boldsymbol{y}_j) - \mu_j)}{\epsilon}\right)}{\sum_j \exp\left(\frac{-(c(\boldsymbol{x}, \boldsymbol{y}_j) - \mu_j)}{\epsilon}\right)} \text{ is soft-min function} \tag{34}$$

Note that compare to (26), there is no hard partition of space $\mathcal{X}$ into Laguerre cells since the smoothed potential function is supported on entire domain. Instead, we assign a point $\boldsymbol{x}$ to one region with probability $[\sigma_j^{\epsilon}(\boldsymbol{x}), j = 1, \ldots, m] \in \Delta_m$, i.e. **soft-assignment**. The gradient is the **difference** between the **discrete measure** b_j at location $\boldsymbol{y}_j$ and $\mathbb{E}_{\alpha}\left[\sigma_j^{\epsilon}(X)\right]$, the expectation of assignment under α.

As stated above, the optimization of discrete β w.r.t. continous $\alpha = \mathcal{N}(m, \sigma^2)$ will result in models such as Gaussian mixtures.

One of important property is that $\nabla_{\boldsymbol{\mu}} \mathcal{E}^{\epsilon}(\boldsymbol{\mu})$ is $1/\epsilon$ Lipschitz, and the Hessian $H(\mathcal{E}^{\epsilon}(\boldsymbol{\mu}))$ is finite and bounded. Similar to logistic regression, $\mathcal{E}^{\epsilon}$ have the properties based on ***self-concordance***. We can solve the problem (31) via **second-order methods** such as *Newton's method*, *quasi-Newton* such as *L-BFGS*.

Since both energy functions in (23) and (31) are the ***expectation*** over α, we can use *stochastic gradient descent* algorithm to solve them.

References

Franz Aurenhammer. Power diagrams: properties, algorithms and applications. *SIAM Journal on Computing*, 16(1):78–96, 1987.

Guillermo Canas and Lorenzo Rosasco. Learning probability measures with respect to optimal transport metrics. *Advances in Neural Information Processing Systems*, 25, 2012.

Gabriel Peyr and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237.

Filippo Santambrogio. *Optimal transport for applied mathematicians*, volume 55. Springer, 2015.