MAT1502 Written Report

Jeffrey Liang

March 2022

1 Introduction

Entropic regularized optimal transport has attracted substantial interest recently after [Cuturi, 2013] demonstrated how it could be used to rapidly compute (approximations to) optimal transport distances between histograms. It also has connections to the Schrödinger problem arising from statistical physics. It modifies the usual Kantorovich formulation of optimal transport by adding an entropic penalty. We aim to give a brief summary of the most important results in this area. Accordingly, we will omit proofs. First we formulate the modified problem and explain some of its basic properties, drawing on [Nutz and Wiesel, 2021]. Then we will discuss its usefulness in data-rich computational problems via Sinkhorn's algorithm, following [Peyré and Cuturi, 2019]. We go on to overview its convergence properties and how to speed it up.

2 The Entropic Regularized Optimal Transport Problem in General

2.1 Formulation

Let (X, μ) and (Y, ν) be Polish probability spaces and $\Gamma(\mu, \nu)$ the set of all couplings of μ and ν, i.e. the probability measures on X $\times$ Y with first marginal μ and second marginal ν. Let c:

$X \times Y \to \mathbb{R}_+$ be continuous and integrable with respect to the product measure. For each $\epsilon > 0$, the entropic regularized optimal transport problem with parameter ϵ is:

$$L_c^\epsilon(\mu, \nu) := \min_{\pi \in \Gamma(\mu,\nu)} \int_{X \times Y} c(x, y) d\pi(x, y) + \epsilon H(\pi|\mu \otimes \nu) \tag{1}$$

where we define

$$H(\pi|\mu \otimes \nu) := \begin{cases} \int_{X \times Y} \log(\frac{d\pi}{d(\mu \otimes \nu)}) d\pi & \text{if } \pi \ll \mu \otimes \nu \\ \infty & \text{otherwise} \end{cases}$$

as the relative entropy of π compared to the product measure. For $\epsilon = 0$, we have the usual Kantorovich optimal transport problem.

This problem may be cast fruitfully as one of probability theory where the objective is to choose the joint distribution of two random variables so as to minimize the expected cost with a preference for distributions with minimal mutual information. Mutual information measures the information one can gain from observing one of the random variables about the other and is minimized at zero precisely when the random variables are independent.

Entropic regularization therefore pushes solutions away from the extreme points of $\Gamma(\mu, \nu)$ where there is very high mutual information. In fact, it first arose in the transportation literature since traffic data does not match the predictions of the usual optimal transport model which suggests very few paths are actually traversed. The solutions to the regularized problem are more realistic since they are more spread out and fuzzy.

2.2 Schrödinger and Kantorovich Potentials

There is a unique minimizer $\pi_\epsilon \in \Gamma(\mu, \nu)$ to (1) due to strict convexity with density

$$\frac{d\pi_\epsilon}{d(\mu \otimes \nu)}(x, y) = \exp\left(\frac{f_\epsilon(x) + g_\epsilon(y) - c(x, y)}{\epsilon}\right) \tag{2}$$

for two measurable functions $f_\epsilon : X \to \mathbb{R}$ and $g_\epsilon : Y \to \mathbb{R}$. This follows from the theory of Schrödinger bridges which we omit for brevity. These functions are called the Schrödinger potentials.

They are unique except that a constant may be added to one and subtracted from the other. Since c is integrable, it follows that $f_\epsilon \in L^1(\mu)$ and $g_\epsilon \in L^1(\nu)$ so we use the symmetric normalization

$$\int_X f_\epsilon(x)d\mu(x) = \int_Y g_\epsilon(y)d\nu(y) \tag{3}$$

in the remainder of Section 2 so that potentials will be unique. They also arise in the dual problem

$$S_c^\epsilon(\mu,\nu) := \sup_{f\in L^1(\mu), g\in L^1(\nu)} \int_X f(x)d\mu(x) + \int_Y g(y)d\nu(y) - \epsilon\int_{X\times Y} e^{\frac{f(x)+g(y)-c(x,y)}{\epsilon}} d(\mu\otimes\nu)(x,y) \tag{4}$$

which is uniquely solved by (f_ϵ, g_ϵ) and where duality holds: $L_c^\epsilon(\mu,\nu) = S_c^\epsilon(\mu,\nu)$. Similarly, the Kantorovich potentials f_0, g_0 arise in the dual problem

$$S_c^0(\mu,\nu) := \sup_{\substack{f_0\in L^1(\mu), g_0\in L^1(\nu),\\ f_0\otimes g_0\leq c}} \int_X f_0(x)d\mu(x) + \int_Y g_0(y)d\nu(y) \tag{5}$$

where $f_0 \otimes g_0(x,y) := f(x) + g(y)$. From standard optimal transport theory, we have duality: $L_c^0(\mu,\nu) = S_c^0(\mu,\nu)$ and the existence of a maximizer (f_0, g_0). We apply the same normalization (3) to these potentials. However, since (5) is not strictly convex, there may still exist multiple pairs of potentials, for example if both marginals are discrete. However, uniqueness is known for many cases of interest, in particular if c is differentiable and at least one marginal support is connected.

2.3 Convergence to the Standard Problem

It is natural to ask whether L_c^ϵ, π_ϵ and (f_ϵ, g_ϵ) converge to their counterparts in the standard Kantorovich formulation as $\epsilon \to 0$. On the primal side, since $\Gamma(\mu,\nu)$ is weakly compact, π_ϵ has cluster points as $\epsilon \to 0$. In fact, the cluster points are solutions to the Kantorovich problem so that $L_c^\epsilon \to L_c^0$. If there is a unique π_0 then we will then have $\pi_\epsilon \to \pi_0$.

On the dual side, [Nutz and Wiesel, 2021] prove that (1) there is a subsequence (ϵ_k) such that f_{ϵ_k} and g_{ϵ_k} converge in $L^1(\mu)$ and $L^1(\nu)$ respectively and (2) if $\lim_n f_{\epsilon_n} = f$ μ-a.e. and $\lim_n g_{\epsilon_n} = g$ ν-a.e. then (f, g) are Kantorovich potentials and convergence also holds in L^1. If the Kantorovich potentials are unique then the whole sequence converges to them in L^1.

3 Applications to Data Science via Sinkhorn's Algorithm

3.1 Motivation

Computational optimal transport has found many uses in a wide variety of fields such as logistics, data sciences, economics and statistics. In these cases, the support of μ and ν are finite. For example, in data science, it is common to create a histogram of features of the data. For a piece of text, one could create a histogram of words showing the relative frequency of each word in the text. Optimal transport can then be leveraged to flexibly compute the distance between any two histograms, parametrized by a pre-defined cost or distance between two words, and thus obtain a measure of similarity between two texts. Another example would be using optimal transport to animate or interpolate between two images.

However, traditional algorithms scaled poorly. If no restrictions were placed on the cost, they had supercubic complexity in the size of the supports. On the other hand, the entropic regularized problem can be approximated using Sinkhorn's algorithm, which is simple and particularly amenable to rapid computation. Each iteration consists of a matrix-vector product and when a large group of histograms share the same support, it can be vectorized and a matrix-matrix product computed with a significant performance improvement. The algorithm converges linearly and in some cases, an $n \times m$ cost matrix does not need to be stored. Finally, the approximate distance that results is smooth with respect to the histogram weights and the location of the Dirac masses and can be differentiated using automatic differentiation, which are desirable properties for machine learning applications.

3.2 Setup

In these cases, the support of μ and ν are finite. The past formulation then simplifies considerably. We may then view μ and ν as probability vectors: non-negative k-dimensional vectors whose elements sum to 1, where k is the cardinality of its support. We will refer to μ as $\mathbf{a}$, a $|\text{supp}(\mu)| = n$-dimensional probability vector, and ν will be $\mathbf{b}$, a $|\text{supp}(\nu)| = m$-dimensional probability vector. The cost function is then just a n x m cost matrix $(\mathbf{C}_{ij})$ with non negative entries that represent the cost associated with mapping the i^{th} mass to the j^{th} target and couplings are the set $U(\mathbf{a}, \mathbf{b})$ of n x m nonnegative matrices with row sum $\mathbf{a}$ and column sum $\mathbf{b}$. The entropic regularized problem is then

$$L_c^{\epsilon}(\mathbf{a}, \mathbf{b}) := \min_{\mathbf{P} \in U(\mathbf{a}, \mathbf{b})} \sum_{i,j} (\mathbf{C}_{ij}\mathbf{P}_{ij} + \epsilon \mathbf{P}_{ij}(\log(\mathbf{P}_{ij}) - 1)) \tag{6}$$

with a unique minimizer $\mathbf{P}_{\epsilon}$. (From now on, when a function of a scalar argument is applied to a matrix, it is meant to be applied entrywise.) In this setting, we have that as $\epsilon \to 0$, $\mathbf{P}_{\epsilon}$ converges to the Kantorovich solution with maximal entropy.

3.3 Sinkhorn's Algorithm

The analogue to (2) is that

$$\mathbf{P}_{ij} = \mathbf{u}_i \mathbf{K}_{ij} \mathbf{v}_j \tag{7}$$

for two scaling vectors $\mathbf{u} \in \mathbb{R}_+^n, \mathbf{v} \in \mathbb{R}_+^m$ and where the kernel matrix is defined as $\mathbf{K}_{ij} := e^{-C_{ij}/\epsilon}$. (We suppress the epsilons for convenience.)

Since $\mathbf{P} \in U(\mathbf{a}, \mathbf{b})$ and rewriting in matrix form, $\mathbf{u}, \mathbf{v}$ must satisfy $diag(\mathbf{u})\mathbf{K}diag(\mathbf{v})\mathbf{1}_m = \mathbf{a}$ and $diag(\mathbf{v})\mathbf{K}^T diag(\mathbf{u})\mathbf{1}_n = \mathbf{b}$. Since $diag(\mathbf{v})\mathbf{1}_m = \mathbf{v}$, this simplifies to

$$\mathbf{u} \odot (\mathbf{K}\mathbf{v}) = \mathbf{a} \qquad \mathbf{v} \odot (\mathbf{K}^T\mathbf{u}) = \mathbf{b} \tag{8}$$

where $\odot$ is entrywise multiplication of vectors. This is known in numerical analysis as the matrix scaling problem. Intuitively we can hope to approximate the solution by altering $\mathbf{u}$ so that it satisfies the first equation of (8) then setting $\mathbf{v}$ so that it satisfies the second and repeating. Sinkhorn's algorithm is precisely this procedure

$$\mathbf{u}^{(l+1)} := \frac{\mathbf{a}}{\mathbf{K}\mathbf{v}^{(l)}} \qquad \mathbf{v}^{(l+1)} := \frac{\mathbf{b}}{\mathbf{K}^T\mathbf{u}^{(l+1)}} \tag{9}$$

initialized with an arbitrary positive vector $\mathbf{v}^{(0)} = 1_m$. Here division between vectors is entrywise. This algorithm in fact converges to the optimal coupling $\mathbf{P}_\epsilon = diag(\mathbf{u})\mathbf{K}diag(\mathbf{v})$ regardless of the initialization vector (though $\mathbf{u}, \mathbf{v}$ may differ because they are only unique up to normalization).

Despite the convergence to the Kantorovich solution with maximal entropy as $\epsilon \to 0$, some issues arise when ϵ becomes very small. Firstly, there are numerical underflow issues. Namely, entries of the kernel matrix $\mathbf{K}$ may become too small to store as positive numbers. This can be partially fixed by performing computations in the log domain. In other words, instead of working with the scaling vectors (**u**,**v**), we work with the Schrödinger potentials $(\epsilon \log(\mathbf{u}), \epsilon \log(\mathbf{v}))$. It is straightforward to work out how Sinkhorn looks in this context and will also be spelled out later in (16). Additionally, the algorithm converges more slowly, especially for costs $\mathbf{C}$ which are close to random. We will discuss this further in the next subsection.

3.4 Convergence Properties of Sinkhorn

To investigate the convergence of Sinkhorn, it is useful to introduce the notion of the Hilbert metric on the projective cone $\mathbb{R}^n_{+,*}$, positive vectors with proportional vectors identified, defined as:

$$d_h(\mathbf{u}, \mathbf{u}') := \log \max_{i,j} \frac{\mathbf{u}_i \mathbf{u}'_j}{\mathbf{u}_j \mathbf{u}'_i} \tag{10}$$

which turns $\mathbb{R}^n_{+,*}$ into a complete metric space. By performing a logarithmic change of variables, we find that this metric is isometric to the variation seminorm:

$$d_h(\mathbf{u}, \mathbf{u}') = || \log \mathbf{u} - \log \mathbf{u}' ||_{var} \tag{11}$$

$$\text{where } ||\mathbf{f}||_{var} := \max_i \mathbf{f}_i - \min_i \mathbf{f}_i \tag{12}$$

The variation seminorm is closely related to the ℓ^∞ norm as we always have $||\mathbf{f}||_{var} \leq 2||\mathbf{f}||_\infty$ and if we impose that $\mathbf{f}_i = 0$ for a fixed i, $|| \cdot ||_{var}$ is a norm and $||\mathbf{f}||_\infty \leq ||\mathbf{f}||_{var}$. These are particularly handy in this case as dual variables $\mathbf{f} = \epsilon \log(\mathbf{u})$ are defined up to additive constants so that we may impose precisely this condition. [Birkhoff, 1957] and [Samelson, 1957] independently introduced the

Hilbert metric and proved the following fundamental theorem, showing that a positive matrix is a contraction on the cone of positive vectors.

Theorem 1 *Let $\boldsymbol{K} \in \mathbb{R}_{+,*}^{p\times r}$ and $(\boldsymbol{v}, \boldsymbol{v}') \in (\mathbb{R}_{+,*}^{r})^2$. Then we have:*

$$d_h(\mathbf{Kv}, \mathbf{Kv}') \leq \lambda(\mathbf{K}) d_h(\mathbf{v}, \mathbf{v}'), \quad \text{where} \quad \begin{cases} \lambda(\mathbf{K}) := \frac{\sqrt{\eta(\mathbf{K})}-1}{\sqrt{\eta(\mathbf{K})}+1} < 1, \\ \eta(\mathbf{K}) := \max_{i,j,k,l} \frac{\mathbf{K}_{i,k}\mathbf{K}_{j,l}}{\mathbf{K}_{j,k}\mathbf{K}_{i,l}} \end{cases}$$

[Franklin and Lorenz, 1989] used this to prove the following theorem which demonstrates the linear convergence of Sinkhorn:

Theorem 2 *In Sinkhorn's algorithm, $(\boldsymbol{u}^{(l)}, \boldsymbol{v}^{(l)}) \to (\boldsymbol{u}^*, \boldsymbol{v}^*)$ and*

$$d_h(\boldsymbol{u}^{(l)}, \boldsymbol{u}^*) = O(\lambda(\boldsymbol{K})^{2l}), \qquad d_h(\boldsymbol{v}^{(l)}, \boldsymbol{v}^*) = O(\lambda(\boldsymbol{K})^{2l}) \tag{13}$$

$$d_h(\boldsymbol{u}^{(l)}, \boldsymbol{u}^*) \leq \frac{d_h(\boldsymbol{P}^{(l)} 1_m, \boldsymbol{a})}{1-\lambda(\boldsymbol{K})^2}, \qquad d_h(\boldsymbol{v}^{(l)}, \boldsymbol{v}^*) \leq \frac{d_h(\boldsymbol{P}^{(l),T} 1_n, \boldsymbol{b})}{1-\lambda(\boldsymbol{K})^2} \tag{14}$$

Here, $(\boldsymbol{u}^*, \boldsymbol{v}^*)$ satisfy $diag(\boldsymbol{u})\boldsymbol{K}diag(\boldsymbol{v}) = \boldsymbol{P}^*$ is the unique minimizer to the entropic regularized problem. We also denote $\boldsymbol{P}^{(l)} := diag(\boldsymbol{u}^{(l)})\boldsymbol{K}diag(\boldsymbol{v}^{(l)})$. Finally we also have

$$|| \log(\boldsymbol{P}^{(l)}) - \log(\boldsymbol{P}^*)||_\infty \leq d_h(\boldsymbol{u}^{(l)}, \boldsymbol{u}^*) + d_h(\boldsymbol{v}^{(l)}, \boldsymbol{v}^*) \tag{15}$$

The bounds (14) show that one can use the amount of the marginal constraint violation to monitor the convergence. By virtue of (11), these bounds give a linear rate for the dual variables $(\mathbf{f}^{(l)}, \mathbf{g}^{(l)}) := (\epsilon \log(\mathbf{u}^{(l)}), \epsilon \log(\mathbf{v}^{(l)}))$ for the variation norm. Also note that the convergence rate degrades as $\epsilon \to 0$. It is easy to see that $\eta(\mathbf{K}^{\epsilon_1}) = \eta(\mathbf{K}^{\epsilon_2})^{\epsilon_2/\epsilon_1}$ and that $\eta(\mathbf{K}) \geq 1$. Therefore, as $\epsilon \to 0$, $\eta(\mathbf{K}^\epsilon) \to \infty$ and $\lambda(\mathbf{K}^\epsilon) \to 1$ for most costs $\mathbf{C}$.

This global linear rate can be quite pessimistic for nice scenarios like if $X = Y = \mathbb{R}^n$ and there is a Monge map when $\epsilon = 0$. In contrast, it is much sharper when the cost $\mathbf{C}$ is close to random, where the rate degrades exponentially in ϵ, $1 - \lambda(\mathbf{K}) \sim e^{-1/\epsilon}$. For a finer asymptotic analysis, one can study the local convergence rate. An iteration of Sinkhorn can be written as the application of a fixed-point map $\mathbf{f}^{(l+1)} = \Phi(\mathbf{f}^{(l)})$ where

$$\Phi := \Phi_2 \circ \Phi_1 \qquad \text{with the definitions} \begin{cases} \Phi_1(\mathbf{f}) := \epsilon(\log(\mathbf{b}) - \log(\mathbf{K}^T(e^{\mathbf{f}/\epsilon}))) \\ \Phi_2(\mathbf{g}) := \epsilon(\log(\mathbf{a}) - \log(\mathbf{K}(e^{\mathbf{g}/\epsilon}))) \end{cases} \tag{16}$$

For Schrödinger potentials $(\mathbf{f}, \mathbf{g})$, the Jacobian of this map is:

$$\partial\Phi(\mathbf{f}) = diag(\mathbf{a})^{-1} \odot \mathbf{P} \odot diag(\mathbf{b})^{-1} \odot \mathbf{P}^T \tag{17}$$

This Jacobian is a positive matrix with $\partial\Phi(\mathbf{f})1_n = 1_n$ so by the Perron-Frobenius theorem, it has a single dominant eigenvector 1_n with eigenvalue 1. Since $\mathbf{f}$ is defined up to a constant, the second eigenvalue $1 - \kappa < 1$ controls the local linear rate of convergence. So for large enough l,

$$||\mathbf{f}^{(l)} - \mathbf{f}|| = O((1 - \kappa)^l) \tag{18}$$

For nice scenarios like when there is a smooth Monge map when $\epsilon = 0$, $\kappa \sim \epsilon$.

Assuming for simplicity that $n = m$, [Altschuler et al., 2017] showed that by choosing $\epsilon = \frac{4\log(n)}{\tau}$, $O(||\mathbf{C}||_\infty^3 \log(n)\tau^{-3})$ Sinkhorn iterations (along with a rounding step to compute a valid coupling) are enough to compute a solution to the unregularized optimal transport problem with cost at most τ more than the optimum. Then in terms of operations, $O(n^2 \log(n)\tau^{-3})$ operations are needed to compute a τ-approximate solution.

3.5 Speeding Up Sinkhorn Iterations

The main bottleneck in Sinkhorn's algorithm are the matrix-vector products with complexity $O(mn)$ in general. However, there are many ways this can be eased significantly. We will mention a few of the most important.

First of all, in the scenario where solutions are sought for multiple problems where only the probability vectors $(\mathbf{a}_i, \mathbf{b}_i)$ vary, the algorithm may be vectorized easily. If we set $\mathbf{A} := (\mathbf{a}_1, ..., \mathbf{a}_N)$

and $\mathbf{B} := (\mathbf{b}_1, ..., \mathbf{b}_N)$, then we can carry out their Sinkhorn iterations in parallel:

$$\mathbf{U}^{(l+1)} := \frac{\mathbf{A}}{\mathbf{K}\mathbf{V}^{(l)}} \qquad \mathbf{V}^{(l+1)} := \frac{\mathbf{B}}{\mathbf{K}^T\mathbf{U}^{(l+1)}} \tag{19}$$

with an arbitrary initialization $\mathbf{V}^0 = 1_{m\times N}$. This consists of matrix-matrix products and can thus be very efficiently executed on GPUs.

A special case where the complexity can be reduced substantially is when each index i, j in the cost matrix can be described as a d-tuple in the Cartesian product of d finite sets,

$$i = (i_k)_{k=1}^d, j = (j_k)_{k=1}^d \in [[n_1]] \times ... \times [[n_d]] \tag{20}$$

and the cost is additive across these subindices. In other words, there exist d matrices $\mathbf{C}^1, ..., \mathbf{C}^d$, each of size $n_1 \times n_1, ..., n_d \times n_d$, such that

$$\mathbf{C}_{ij} = \sum_{k=1}^{d} \mathbf{C}^k_{i_k, j_k}, \tag{21}$$

then because of the definition of $\mathbf{K}$ it has a separable structure:

$$\mathbf{K}_{ij} = \prod_{k=1}^{d} \mathbf{K}^k_{i_k, j_k}. \tag{22}$$

This structure allows for a very fast evaluation of $\mathbf{Ku}$. Instead of instantiating $\mathbf{K}$ as a $n \times n$ matrix which would usually be infeasible because $n = \prod_{k=1}^d n_k$ is exponential in d, we can apply $\mathbf{K}^k$ along the appropriate slice of $\mathbf{u}$. If $n = m$, the complexity of a Sinkhorn iteration drops from $O(n^2)$ to $O(n^{1+1/d})$.

Let us illustrate with an important example. Suppose $X = Y = [0,1]^d$, the cost is the pth power of the p-norm, and the space is discretized with a regular grid so that we only consider points $x_i = (i_1/n_1, ..., i_d/n_d)$ for $i = (i_1, ..., i_d) \in [[n_1]] \times ... \times [[n_d]]$. It is now very efficient to compute $\mathbf{Ku}$ by applying the $n_k \times n_k$ matrices $\mathbf{K}^k_{rs} = \exp(-|\frac{r-s}{n_k}|^p/\epsilon)$ to $\mathbf{u}$ reshaped as a tensor whose first dimension has been permuted so as to match the kth set of indices. For example, if

$d = 2$ and $p = 2$, histograms $\mathbf{a}$ and scaling vectors $\mathbf{u}$ can be instantiated as $n_1 \times n_2$ matrices with the first dimension permuted to match the first set of indices $(j_1)_{1 \leq j \leq m}$. We write $\mathbf{U}$ instead of $\mathbf{u}$ to underline that it is shaped like a $n_1 \times n_2$ matrix instead of a vector of length $n_1 n_2$. Then we can compute $\mathbf{Ku}$ which needs $(n_1 n_2)^2$ operations naively by applying the 1-D convolutions one at a time,

$$(\mathbf{K}^2(\mathbf{K}^1\mathbf{U})^T)^T = \mathbf{K}^1\mathbf{U}\mathbf{K}^2 \tag{23}$$

in only $n_1^2 n_2 + n_1 n_2^2$ operations. This procedure generalizes to larger d and is very efficient on GPUs.

3.6 Regularized Optimal Transport Distances

One can view the regularized distances as improvements over the standard Optimal Transport distances instead of viewing them as introducing approximation error. For example, they are smooth and convex which makes them well suited for use as loss functions in machine learning.

To be specific, $L_c^\epsilon(\mathbf{a}, \mathbf{b})$ is jointly convex in $\mathbf{a}$ and $\mathbf{b}$. When $\epsilon > 0$, its gradient is

$$\nabla L_c^\epsilon(\mathbf{a}, \mathbf{b}) = [\mathbf{f}^*, \mathbf{g}^*], \tag{24}$$

where $\mathbf{f}^*, \mathbf{g}^*$ are the Schrödinger potentials normalized so their coordinates sum to zero.

Sinkhorn gives both a lower and upper bound on the regularized distances $L_c^\epsilon(\mathbf{a}, \mathbf{b})$ by virtue of the primal and dual characterizations. Given $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ computed by Sinkhorn, let $\hat{\mathbf{P}}$ represent the matrix computed by rounding $diag(\mathbf{u})\mathbf{K}diag(\mathbf{v})$ so that it is a valid coupling. Then we have the following bounds:

$$\epsilon(\log(\hat{\mathbf{u}}) \cdot \mathbf{a} + \log(\hat{\mathbf{v}}) \cdot \mathbf{b} - \hat{\mathbf{u}}\mathbf{K}\hat{\mathbf{v}}) \leq L_c^\epsilon(\mathbf{a}, \mathbf{b}) \leq \sum_{i,j}(\mathbf{C}_{ij}\hat{\mathbf{P}}_{ij} + \epsilon\hat{\mathbf{P}}_{ij}(\log(\hat{\mathbf{P}}_{ij}) - 1)) \tag{25}$$

References

[Altschuler et al., 2017] Altschuler, J., Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *CoRR*, abs/1705.09634.

[Birkhoff, 1957] Birkhoff, G. (1957). Extensions of jentzsch's theorem. *Transactions of the American Mathematical Society*, 85(1):219–227.

[Cuturi, 2013] Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transportation distances.

[Franklin and Lorenz, 1989] Franklin, J. and Lorenz, J. (1989). On the scaling of multidimensional matrices. *Linear Algebra and its applications*, 114:717–735.

[Nutz, 2022] Nutz, M. (2022). Introduction to entropic optimal transport. `https://www.math.columbia.edu/~mnutz/docs/EOT_lecture_notes.pdf`. [Online, accessed 20-Mar-2022].

[Nutz and Wiesel, 2021] Nutz, M. and Wiesel, J. (2021). Entropic optimal transport: Convergence of potentials. *Probability Theory and Related Fields*, pages 1–24.

[Peyré and Cuturi, 2019] Peyré, G. and Cuturi, M. (2019). Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607.

[Samelson, 1957] Samelson, H. (1957). On the perron-frobenius theorem. *Michigan Mathematical Journal*, 4(1):57–59.