Stochastic Algorithm for Optimal Transport

Zehui XUAN

Master IASD

April - September 2022

Advisor :
Olivier Wintenberger
Antoine Godichon-Baggioni

Contents

Abstract

Optimal Transport has an increasingly wide range of applications nowadays, where the squared Wasserstein distance is a useful quantity. However it is difficult to be estimated due to a large computational burden and the errors caused by regularization. In this paper, we use Sinkhorn divergence as the estimator and introduce a regularization level ε that decreases with iteration, in order to facilitate the computation in the beginning of iterations and reduce the error in the later iterations. Our theoretical analysis shows that using the Stochastic Average Gradient (SAG) algorithm cannot avoid a regret bound of higher order of T, where T is the number of iteration, while using the Averaged SGD (ASGD) algorithm has a better theoretical regret bound. However, our experimental results show that SAG also performs well under certain ε settings.

Chapter 1

Introduction

Nowadays, Optimal Transport (OT) are increasingly being applied in machine learning problems such as collaborative filtering [LCCC21], sorting [CTV19], matching [LYZZ19], generative model [GPC17] [TBGS19]. The OT problem starts with finding an optimal transportation plan to transport small stones from some locations to others with minimum cost. Following continuous development and expansion, OT theory can now be utilized to measure how much difference is between two probability measures $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$.

When μ and ν are discrete measures, the distance between them can be expressed by the Kantorovich formulation [Kan42], which can be seen as a Linear Program. Its optimal solution between two measures of N-bins can be found by simplex algorithm, which have a worst-case complexity of exponential time, while the time complexity of interior-point algorithms is $O(N^3 \ln N)$ [PW09]. In [Cut13], the authors proposed Entropy regularized OT, which transforms the original OT problem into a strict convex problem that can be solved by the Sinkhorn algorithm with a time complexity of $O(N^2 \ln N/\lambda^2)$ to achieve a λ-accuracy [DGK18], greatly speeding up the computation. In [GCPB16], the authors used the stochastic algorithms for computing large-scale OT. This is based on two ideas: the maximization of expectations and the smoothness of dual problem of regularized OT. In the discrete case, the authors used the Stochastic Average Gradient (SAG) algorithm (1) to find an optimum for the semi-dual problem of regularized OT, which has a convergence rate of $O(1/k)$ where k is the number of iteration.

We are more concerned with the semi-discrete case, where μ is an arbitrary measure and ν is a discrete measure. One application of this case is that ν is an empirical measure of some sample we have and μ is a theoretical measure and we want to compute the distance between them. It is more complicated than the previous case, because if we want to utilize the previously mentioned algorithms, we first need to discretize μ into $\hat{\mu}_n$, a discrete measure. When n is not large enough, the error generated by the discretization is not negligible. If n is particularly large, it will increase the computational burden. Still in [GCPB16], the authors used the Averaged SGD algorithm (2) for the same objective in semi-discrete case, which has a convergence rate of $O(1/\sqrt{k})$ where k is the number of iteration. The merit of this algorithm is that it does not require μ to be discrete. It will sample from μ at the beginning of each iteration, and after a large number of iterations, the final output

naturally takes advantage of these previous samples. In [BBGS22], the authors introduce a stochastic Gauss-Newton (SGN) algorithm to estimate the regularized OT cost for the semi-discrete case. The advantage of this algorithm is adaptive so that it does not require important hyperparameter tuning.

The value we are more interested in is the non-regularized OT cost, in particular the Squared Wasserstein Distance. In [DD20], in order to solve the optimal transport problem between two Gaussian mixture distributions, the authors restricted the set of optimal transport plan to a set of mixed Gaussian models, which transform the original problem into a computationally simple problem. The obtained transport cost is Mixture Wasserstein (MW_2^2), a variant of squared Wasserstein distance, which is upper bounded by the solution of the original problem plus a term that depends only on the parameters of the Gaussian mixture models. In [CRL+20], the authors proposed empirical Sinkhorn Divergence as an estimator of Squared Wasserstein Distance between two continuous measures and solved it with the Sinkhorn algorithm. Although the error caused by discretization cannot be avoided in this method, this estimator overcomes the plug-in estimator $W_2^2(\hat{\mu}_n, \hat{\nu}_n)$ computational burden problem and reduces the bias problem of the regularized OT.

Our contributions in this paper are listed in the following.

- We used Empirical Sinkhorn Divergence to estimate the squared Wasserstein distance and then solved it by using SAG algorithm (1). We also performed the regret analysis under this setting.
- Based on the previous point, we replaced the hyperparameter ε with a ε_t that decreases with iteration, and then carried out a regret analysis.

We list the symbols that will be used next sections in the paper here.

- $\mathcal{X}$, $\mathcal{Y}$: two metric spaces
- $\mathcal{P}(\mathcal{X})$: set of probability measures on $\mathcal{X}$.
- $\mu \in \mathcal{P}(\mathcal{X}), \nu \in \mathcal{P}(\mathcal{Y})$: two probability measures
- $\Pi(\mu, \nu)$: set of joint probability measures on $\mathcal{X} \times \mathcal{Y}$, with marginal $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$.
- $\mu \otimes \nu$: product measure of μ and ν.
- $\mathcal{C}(\mathcal{X})$: space of continuous function on $\mathcal{X}$.
- Σ_n: probability simplex with n bins.

This paper is organized as follows: In section 2, we provide a reminder for the optimal transport theory and involved stochastic algorithms. Section 3 introduces our analysis of the regret bound in several different settings. In section 4, we present the the procedure and results of the numerical experiments before the conclusion.

Chapter 2

Preliminaries

2.1 Optimal Transport

2.1.1 Kantorovich formulation and Squared Wasserstein Distance

The Kantorovich formulation [Kan42] of the optimal transport problem between μ and ν can be formulated as following convex minimization problem

$$\forall(\mu,\nu)\in\mathcal{P}(\mathcal{X})\times\mathcal{P}(\mathcal{Y}), W(\mu,\nu) \overset{\text{def.}}{=} \min_{\pi\in\Pi(\mu,\nu)}\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) \tag{2.1}$$

where $c(x,y)$ is the cost of transporting a unit of mass from x to y. The quantity $W(\mu,\nu)$ is the optimal total transportation cost.

When $c(x,y)=\|x-y\|^2$, the total cost $W(\mu,\nu)$ in (2.1) is called Squared Wasserstein Distance [CRL+20] and we denote it by $W_2^2(\mu,\nu)$.

$$W_2^2(\mu,\nu) \overset{\text{def.}}{=} \min_{\pi\in\Pi(\mu,\nu)}\int_{\mathcal{X}\times\mathcal{Y}} \|x-y\|^2\mathrm{d}\pi(x,y) \tag{2.2}$$

When $\mu=\mathcal{N}(m_0,\Sigma_0)$ and $\nu=\mathcal{N}(m_1,\Sigma_1)$ are two Gaussian distribution, Squared Wasserstein Distance has an explicit expression [DD20].

$$W_2^2(\mu,\nu)=\|m_0-m_1\|^2+\mathrm{tr}\left(\Sigma_0+\Sigma_1-2\left(\Sigma_0^{\frac{1}{2}}\Sigma_1\Sigma_0^{\frac{1}{2}}\right)^{\frac{1}{2}}\right) \tag{2.3}$$

However, when μ and ν are not the Gaussian distribution, we want to find a way to approximate this quantity, especially by using the stochastic algorithms, which is our main goal for this internship.

2.1.2 Regularized OT and its dual and semi-dual formulation

The regularized OT problem and its cost were proposed in [Cut13] in order to alleviate the computational cost in the original problem (2.1)

$$W_\varepsilon(\mu,\nu) \overset{\text{def.}}{=} \min_{\pi\in\Pi(\mu,\nu)}\int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y)+\varepsilon\mathrm{KL}(\pi\mid\mu\otimes\nu) \tag{2.4}$$

where $\varepsilon \geq 0$ is the regularization parameter and the regularization term $\mathrm{KL}(\pi \mid \mu \otimes \nu)$ is the Kullback-Leibler divergence.

Dual formulation of regularized OT problem [GCPB16] is in a form of the following concave maximization problem

$$W_\varepsilon(\mu,\nu) = \max_{u\in\mathcal{C}(\mathcal{X}),v\in\mathcal{C}(\mathcal{Y})} F_\varepsilon(u,v) \overset{\text{def.}}{=} \int_{\mathcal{X}} u(x)\mathrm{d}\mu(x) + \int_{\mathcal{Y}} v(y)\mathrm{d}\nu(y) - \iota^\varepsilon_{U_c}(u,v) \quad (2.5)$$

where U_c is a constraint set defined as following

$$U_c \overset{\text{def.}}{=} \{(u,v) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}); \forall (x,y) \in \mathcal{X}\times\mathcal{Y}, u(x)+v(y) \leq c(x,y)\} \quad (2.6)$$

and $\iota^\varepsilon_{U_c}$ is a smoothed approximation of its indicator function

$$\iota^\varepsilon_{U_c}(u,v) \overset{\text{def.}}{=} \begin{cases} \iota_{U_c}(u,v) \text{ if } \varepsilon = 0, \\ \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}\right) \mathrm{d}\mu(x)\mathrm{d}\nu(y) \text{ if } \quad \varepsilon > 0 \end{cases} \quad (2.7)$$

The semi-dual formulation of this regularized OT problem [GCPB16] is also in a form of the following concave maximization problem

$$W_\varepsilon(\mu,\nu) = \max_{v\in\mathcal{C}(\mathcal{Y})} H_\varepsilon(v) \overset{\text{def.}}{=} \int_{\mathcal{X}} v^{c,\varepsilon}(x)\mathrm{d}\mu(x) + \int_{\mathcal{Y}} v(y)\mathrm{d}\nu(y) - \varepsilon \quad (2.8)$$

where $v^{c,\varepsilon}(x)$ is the c-transform and the approximation for any $v \in \mathcal{C}(\mathcal{Y})$:

$$\forall x \in \mathcal{X}, \quad v^{c,\varepsilon}(x) \overset{\text{def.}}{=} \begin{cases} \min_{y\in\mathcal{Y}} c(x,y) - v(y) & \text{if} \quad \varepsilon = 0 \\ -\varepsilon \ln\left(\int_{\mathcal{Y}} \exp\left(\frac{v(y)-c(x,y)}{\varepsilon}\right) \mathrm{d}\nu(y)\right) & \text{if} \quad \varepsilon > 0 \end{cases} \quad (2.9)$$

We also present the semi-dual formulation in the case where $\mu = \sum_{i=1}^I \boldsymbol{\mu}_i \delta_{x_i}$ and $\nu = \sum_{j=1}^J \boldsymbol{\nu}_j \delta_{y_j}$ are both discrete, since it is crucial to the algorithm we used in this work. More precisely, we have

$$W_\varepsilon(\mu,\nu) = \max_{\mathbf{v}\in\mathbb{R}^J} \bar{H}_\varepsilon(\mathbf{v}) = \sum_{i=1}^I \bar{h}_\varepsilon(x_i,\mathbf{v})\boldsymbol{\mu}_i, \quad (2.10)$$

where

$$\bar{h}_\varepsilon(x,\mathbf{v}) = \sum_{j=1}^J \mathbf{v}_j\boldsymbol{\nu}_j + \begin{cases} -\varepsilon \ln\left(\sum_{j=1}^J \exp\left(\frac{\mathbf{v}_j - c(x,y_j)}{\varepsilon}\right)\boldsymbol{\nu}_j\right) - \varepsilon & \text{if } \varepsilon > 0, \\ \min_j \left(c\left(x,y_j\right) - \mathbf{v}_j\right) & \text{if } \varepsilon = 0, \end{cases} \quad (2.11)$$

and its gradient

$$\nabla_v \bar{h}_\varepsilon(x,\mathbf{v}) = \boldsymbol{\nu} - \chi^\varepsilon_{(c(x,y_\ell)-\mathbf{v}_\ell)_\ell}, \quad \text{where} \quad \forall \varepsilon > 0, (\chi^\varepsilon_r)_j \overset{\text{def.}}{=} e^{-\frac{r_j}{\varepsilon}}\boldsymbol{\nu}_j \left(\sum_\ell e^{-\frac{r_\ell}{\varepsilon}}\boldsymbol{\nu}_\ell\right)^{-1}. \quad (2.12)$$

In the case of semi-discrete where μ can be arbitrary measure and $\nu = \sum_{j=1}^J \boldsymbol{\nu}_j \delta_{y_j}$ is a discrete measure, the semi-dual problem is in a form of maximization of expectation as follows.

$$W_\varepsilon(\mu,\nu) = \max_{\mathbf{v}\in\mathbb{R}^J} \mathbf{E}_X\left[\bar{h}_\varepsilon(X,\mathbf{v})\right] \quad (2.13)$$

where $X \sim \mu$.

2.1.3 Sinkhorn Divergence

In [CRL+20], the authors proposed an estimator $\hat{S}_{\varepsilon,n} = S_\varepsilon(\hat{\mu}_n, \hat{\nu}_n)$ to approximate $W_2^2(\mu, \nu)$, where S_ε is the Sinkhorn Divergence:

$$S_\varepsilon(\mu, \nu) \stackrel{\text{def.}}{=} W_\varepsilon(\mu, \nu) - \frac{1}{2}\left(W_\varepsilon(\mu, \mu) + W_\varepsilon(\nu, \nu)\right) \tag{2.14}$$

Sinkhorn Divergence adds a debiasing term compared to $W_\varepsilon(\mu, \nu)$, which allows it to be zero when $\mu = \nu$, a property that $W_\varepsilon(\mu, \nu)$ does not possess. In [CRL+20] the authors also showed that it has an error of $O(\varepsilon^2)$ relative to $W_2^2(\mu, \nu)$, which is better than the error $O(\varepsilon \ln(1/\varepsilon))$ of $W_\varepsilon(\mu, \nu)$.

2.2 Stochastic algorithms

2.2.1 Stochastic Average Gradient (SAG)

Stochastic Average Gradient (SAG) algorithm was first proposed in [SRB13]. Considering the following optimization problem,

$$\text{minimize}_{x \in \mathbb{R}^p} \quad g(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{2.15}$$

at each iteration, SAG only accesses an index i sampled from $\{1, ..., n\}$ like SGD, but it uses the average of the gradients of all f_i as a proxy of the gradient of g, by keeping memory of all the last updated f_i'. This allows the iteration cost to become relatively lower than Deterministic gradient descent, especially when we have a large number of samples from μ, but has a better proxy of gradient of g than SGD.

The SAG iteration can be described in the following form.

$$x^{k+1} = x^k - \frac{\alpha_k}{n}\sum_{i=1}^{n} y_i^k \tag{2.16}$$

where y_i^k will be updated as $f_i'(x^k)$ if i_k is randomly selected from $\{1, ..., n\}$, otherwise y_i^k will remain the same as the previous iteration.

$$y_i^k = \begin{cases} f_i'\left(x^k\right) & \text{if } i = i_k \\ y_i^{k-1} & \text{otherwise} \end{cases} \tag{2.17}$$

Then SAG was used to solve the discrete OT problem between $\mu = \sum_{i=1}^{I} \boldsymbol{\mu}_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{J} \boldsymbol{\nu}_j \delta_{y_j}$ in [GCPB16], shown in Algorithm (1).

Algorithm 1 SAG for Discrete OT between μ and ν

Input: α
Output: $\mathbf{v}$
$\mathbf{v} \leftarrow \mathbb{0}_J$, $\mathbf{d} \leftarrow \mathbb{0}_J$, $\forall i, \mathbf{g}_i \leftarrow \mathbb{0}_J$
for $k = 1, 2, ...$ **do**
Sample $i \in \{1, 2, ..., I\}$ uniform.
$\mathbf{d} \leftarrow \mathbf{d} - \mathbf{g}_i$
$\mathbf{g}_i \leftarrow \boldsymbol{\mu}_i \nabla_v \bar{h}_\varepsilon(x_i, \mathbf{v})$
$\mathbf{d} \leftarrow \mathbf{d} + \mathbf{g}_i$
$\mathbf{v} \leftarrow \mathbf{v} + \alpha \mathbf{d}$
end for

2.2.2 Averaged SGD (ASGD)

Averaged SGD was first introduced in the article [PJ92]. The intuition of this algorithm is averaging the historical path of the classical SGD which brings more stability. Still considering the problem (2.15), the iteration can be described as follows.

$$\tilde{x}^{k+1} = \tilde{x}^k - \alpha_k f'_{i_k}(\tilde{x}^k), \tag{2.18}$$

$$x^k = \frac{1}{k} \sum_{\ell=1}^{k} \tilde{x}^\ell \tag{2.19}$$

where x_k is the output of the algorithm. Then Averaged SGD was adopted to solve the semi-discrete OT problem between an arbitrary measure μ and a discrete measure $\nu = \sum_{j=1}^{J} \boldsymbol{\nu}_j \delta_{y_j}$ in [GCPB16] which is presented in Algorithm (2).

Algorithm 2 Averaged SGD for Semi-Discrete OT

Input: α
Output: $\mathbf{v}$
$\tilde{\mathbf{v}} \leftarrow \mathbb{0}_J$, $\mathbf{v} \leftarrow \tilde{\mathbf{v}}$
for $k = 1, 2, ...$ **do**
Sample x_k from μ
$\tilde{\mathbf{v}} \leftarrow \tilde{\mathbf{v}} + \frac{\alpha}{\sqrt{k}} \nabla_v \bar{h}_\varepsilon(x_k, \tilde{\mathbf{v}})$
$\mathbf{v} \leftarrow \frac{1}{k} \tilde{\mathbf{v}} + \frac{k-1}{k} \mathbf{v}$
end for

Chapter 3

Regret Analysis

3.1 Regret Analysis for the Sinkhorn Divergence

3.1.1 Notations

The notations used in this subsection are listing here:

- μ: a continuous measure in $\mathcal{P}(\mathbb{R}^d)$. In this analysis, we only consider the case where $d > 1$.
- $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$: discretized μ with sample size n.
- $\nu_J = \sum_{j=1}^{J} \boldsymbol{\nu}_j \delta_{y_j}$: a discrete measure in $\mathcal{P}(\mathbb{R}^d)$ with sample size J, where $\boldsymbol{\nu} \in \Sigma_J$ is a weight vector for ν_J and the weights are random.
- W_2^2: $W_2^2(\mu, \nu_J)$, squared Wasserstein distance between μ and ν_J.
- S_ε: $S_\varepsilon(\mu, \nu_J) = W_\varepsilon(\mu, \nu_J) - \frac{1}{2}(W_\varepsilon(\mu, \mu) + W_\varepsilon(\nu_J, \nu_J))$, Sinkhorn Divergence between μ and ν_J.
- $\hat{S}_{\varepsilon,n}$: $\hat{S}_{\varepsilon,n}(\mu, \nu_J) = S_\varepsilon(\hat{\mu}_n, \nu_J)$, empirical Sinkhorn Divergence between μ and ν_J, in semi-discrete setting.
- $\hat{S}_{\varepsilon,n}^{(t)}$: estimation of $\hat{S}_{\varepsilon,n}$ at iteration t, using SAG algorithm (1).

3.1.2 Introduction

In this subsection, we analyse a bound for $\sum_{t=1}^{T} \left| \mathbf{E}\left[\hat{S}_{\varepsilon,n}^{(t)}\right] - W_2^2 \right|$, where $\hat{S}_{\varepsilon,n}^{(t)}$ is an approximation of the empirical Sinkhorn Divergence using SAG algorithm (1) at iteration t. Two cases are considered in this analysis. In the first case, the regularization level ε and the number of discretization n for μ are fixed in advance. After obtaining the expression of this bound, we try to find an optimal pair of ε and n in functions of T to minimize this bound. In the second case, n is also fixed in advance while we replace the fixed ε in the previous case with ε_t that decreases as the iteration t increases. For example, we can make $\varepsilon_t = \frac{\varepsilon_0}{\sqrt{t}}$ for $t \geq 1$. The intuition

is that in the preceding iterations, ε_t is relatively larger, which reduces the difficulty of approximation. In the later iterations, ε_t decreases, which allows the bias of our objective decreases, thus our objective becomes closer to the original problem. This is a very nice way to balance the trade-off between computational difficulty and bias. Once we get the bound, we also try to find an optimal n as a function of T.

To realize this analysis, we first consider the following decomposition:

$$\left|\mathbf{E}\left[\hat{S}_{\varepsilon,n}^{(t)}\right] - W_2^2\right| \leq \left|\mathbf{E}\left[\hat{S}_{\varepsilon,n}^{(t)}\right] - \hat{S}_{\varepsilon,n}\right| + \left|\hat{S}_{\varepsilon,n} - \mathbf{E}\left[\hat{S}_{\varepsilon,n}\right]\right| + \mathbf{E}\left[\left|\hat{S}_{\varepsilon,n} - S_\varepsilon\right|\right] + \left|S_\varepsilon - W_2^2\right| \tag{3.1}$$

Afterwards, we adapt several pertinent propositions in [CRL+20], [MN19] and [SRB13] to our case (semi-discrete) in order to find the bounds for the components in this decomposition.

3.1.3 A bound for $\left|S_\varepsilon - W_2^2\right|$

In [CRL+20], the author mentioned that using Sinkhorn Divergence $S_\varepsilon(\mu,\nu)$ to approximate $W_2^2(\mu,\nu)$ produces an error of order $O(\varepsilon^2)$, assuming that μ and ν have bounded densities and supports. We can write this error as following.

$$\left|S_\varepsilon(\mu,\nu) - W_2^2(\mu,\nu)\right| \leq O(\varepsilon^2) \tag{3.2}$$

In our case, because ν is discrete, it does not meet the requirement of bounded density. Nevertheless, since we did not find a conclusion about it, we assume that this error holds in the semi-discrete case as well.

3.1.4 A bound for $\mathbf{E}\left[\left|\hat{S}_{\varepsilon,n} - S_\varepsilon\right|\right]$

We obtain the empirical Sinkhorn divergence by discretization of the continuous measure μ in the Sinkhorn divergence, which introduces an error related to the discretization. In [CRL+20], Lemma 5, the authors provided a bound of expectation of the error between empirical Sinkhorn divergence and Sinkhorn divergence in the continuous setting where μ and ν are both continuous measures. We next find this bound in the semi-discrete case by adapting this lemma.

Proposition 1. *Let $(\mu,\nu) \in \mathcal{P}(\mathbb{R}^d) \times \mathcal{P}(\mathbb{R}^d)$ be concentrated on a set of diameter 1, and let $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ be a empirical measure of μ where $x_i, i = 1...n$ are points independently sampled from μ. Let $d' = 2\lfloor d/2 \rfloor$. Then*

$$\mathbf{E}\left[|S_\varepsilon(\hat{\mu}_n,\nu) - S_\varepsilon(\mu,\nu)|\right] \lesssim \left(1 + \varepsilon^{-d'/2}\right) n^{-1/2} \tag{3.3}$$

where $\lesssim$ hides a constant that only depends on d.

We can interpret this theorem by saying that as ε decreases, we need more samples from μ to keep this error from increasing.

Proof. We consider the following decomposition of this error.

$$|S_\varepsilon(\hat{\mu}_n,\nu) - S_\varepsilon(\mu,\nu)| \leq |W_\varepsilon(\hat{\mu}_n,\nu) - W_\varepsilon(\mu,\nu)| + \frac{1}{2}|W_\varepsilon(\hat{\mu}_n,\hat{\mu}_n) - W_\varepsilon(\mu,\mu)| \tag{3.4}$$

The second term of the right hand side has been shown in the proof of lemma 5 in [CRL+20] to have a bound of $\left(1+\varepsilon^{-d'/2}\right) n^{-1/2}$ up to a multiplicative constant.

According to the Proposition 2 in [MN19], we have

$$|W_\varepsilon(\hat{\mu}_n,\nu) - W_\varepsilon(\mu,\nu)| \leq 2\sup_{f\in\mathcal{F}}\left|\int f \mathrm{d}(\hat{\mu}_n-\mu)\right| \tag{3.5}$$

where $\mathcal{F}$ is a set of function that contains all the solution satisfying (2.4), for all pairs of measures $(\mu,\nu)\in\mathcal{P}(\mathbb{R}^d)\times\mathcal{P}(\mathbb{R}^d)$, where the support of the measures are concentrated on a set of diameter 1.

It was shown in the proof of Lemma 5 of [CRL+20] that the expectation of the bound in (3.5) have a bound

$$\mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\int f \mathrm{d}(\hat{\mu}_n-\mu)\right|\right] \lesssim \left(1+\varepsilon^{-d'/2}\right) n^{-1/2} \tag{3.6}$$

where $\lesssim$ hides a constant that only depends on d.

Since the expectation of two components in the right hand side of (3.4) have same statistical bound, we can conclude that the expectation of its left hand side also share this statistical bound.

□

3.1.5 A bound for $\left|\hat{S}_{\varepsilon,n} - \mathbf{E}\left[\hat{S}_{\varepsilon,n}\right]\right|$

The discretization of the Sinkhorn divergence mentioned in the section 3.1.4 is stochastic, therefore we also need to find a bound to describe the concentration of the empirical Sinkhorn divergence. In [CRL+20], Proposition 4, the authors presented this concentration bound in the continuous setting. We also adapt this proposition in the semi-discrete setting.

Proposition 2. *Let $\mu\in\mathcal{P}(\mathbb{R}^d)$ be a continuous measure, and $\nu_J\in\mathcal{P}(\mathbb{R}^d)$ be a discrete measure of size J whose weights are random. Both of them are concentrated on a set of diameter 1. Let $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ be a empirical measure of μ where $x_i, i=1...n$ are points independently sampled from μ. Then*

$$\mathbf{P}\left[\left|\hat{S}_{\varepsilon,n} - \mathbf{E}\left[\hat{S}_{\varepsilon,n}\right]\right| \geq s\right] \leq 2\exp\left(-\frac{s^2}{2\left(\frac{1}{n}+J\right)}\right) \tag{3.7}$$

Proof. As in [CRL+20], Proof of Proposition 12, we prove this bound by considering the Primal problem (2.4). We have the empirical Sinkhorn Divergence $\hat{S}_{\varepsilon,n} = S_\varepsilon(\hat{\mu}_n,\nu_J)$ where $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n\delta_{x_i}$, $\nu_J = \sum_{i=j=1}^J \boldsymbol{\nu}_j\delta_{y_j}$ and $\boldsymbol{\nu}\in\Sigma_J$. Let $\boldsymbol{\mu} = \frac{1}{n}\cdot\mathbf{1}_n$ denote the weight vector for $\hat{\mu}_n$. Let $c\in\mathbb{R}^{n\times J}$ be a cost matrix where its entries $c_{i,j}$ are $\|x_i-y_j\|_2^2$. The primal problem can be written as following.

$$W_\varepsilon(\hat{\mu}_n,\nu_J) = \min_{P\in\mathbb{R}_+^{n\times J}, P\mathbf{1}_J=\boldsymbol{\mu}, P^\top\mathbf{1}_n=\boldsymbol{\nu}} \sum_{i,j} c_{i,j}P_{i,j} + \sum_{i,j} P_{i,j}(\ln(P_{i,j})-1) \tag{3.8}$$

Let P^* be the minimizer of this problem. Then we replace the last point x_n in the sample of μ by a new point x'_n which is also in the same set of diameter 1. Let $\hat{\mu}'_n = \frac{1}{n}(\sum_{i=1}^{n-1} \delta_{x_i} + \delta_{x'_n})$ be the new sample of μ. This causes the changes of value in the last row of the cost matrix, where the largest change of $c_{n,j}, j = 1...J$ is 1. Let c' denote this new cost matrix. Then we have $\left|c_{i,j} - c'_{i,j}\right| = 0$ if $i \leq n-1$ and $\left|c_{i,j} - c'_{i,j}\right| \leq 1$ if $i = n$.

We define $W_\varepsilon(\hat{\mu}'_n, \nu_J)$ as the total transport cost between $\hat{\mu}'_n$ and ν_J by adapting P^* as the transport plan. Since the sum of the last row of P^* is $\frac{1}{n}$, we have the bound of the difference between $W_\varepsilon(\hat{\mu}'_n, \nu_J)$ and $W_\varepsilon(\hat{\mu}_n, \nu_J)$.

$$\begin{aligned} |W_\varepsilon(\hat{\mu}_n, \nu_J) - W_\varepsilon(\hat{\mu}'_n, \nu_J)| &= \left|\sum_{i,j}(c_{i,j} - c'_{i,j}) \cdot P^*_{i,j}\right| \\ &= \left|\sum_{j=1}^{J}(c_{n,j} - c'_{n,j}) \cdot P^*_{n,j}\right| \\ &\leq \frac{1}{n} \end{aligned} \tag{3.9}$$

A similar argument can be shown to obtain a bound for perturbing one sample in ν_J.

$$|W_\varepsilon(\hat{\mu}_n, \nu'_J) - W_\varepsilon(\hat{\mu}_n, \nu_J)| \leq 1 \tag{3.10}$$

Then we consider the bound of replacing one sample in $\hat{\mu}_n$ for Sinkhorn divergence.

$$\begin{aligned} &|S_\varepsilon(\hat{\mu}'_n, \nu_J) - S_\varepsilon(\hat{\mu}_n, \nu_J)| \\ =&\left|(W_\varepsilon(\hat{\mu}'_n, \nu_J) - \frac{1}{2}W_\varepsilon(\hat{\mu}'_n, \hat{\mu}'_n) - \frac{1}{2}W_\varepsilon(\nu_J, \nu_J)) - (W_\varepsilon(\hat{\mu}_n, \nu_J) - \frac{1}{2}W_\varepsilon(\hat{\mu}_n, \hat{\mu}_n) - \frac{1}{2}W_\varepsilon(\nu_J, \nu_J))\right| \\ \leq&|W_\varepsilon(\hat{\mu}'_n, \nu_J) - W_\varepsilon(\hat{\mu}_n, \nu_J)| + \frac{1}{2}|W_\varepsilon(\hat{\mu}'_n, \hat{\mu}'_n) - W_\varepsilon(\hat{\mu}_n, \hat{\mu}_n)| \\ \leq&|W_\varepsilon(\hat{\mu}'_n, \nu_J) - W_\varepsilon(\hat{\mu}_n, \nu_J)| + \frac{1}{2}(|W_\varepsilon(\hat{\mu}'_n, \hat{\mu}'_n) - W_\varepsilon(\hat{\mu}_n, \hat{\mu}'_n)| + |W_\varepsilon(\hat{\mu}_n, \hat{\mu}'_n) - W_\varepsilon(\hat{\mu}_n, \hat{\mu}_n)|) \\ \leq&\frac{1}{n} + \frac{1}{2} \cdot \frac{2}{n} = \frac{2}{n} \end{aligned} \tag{3.11}$$

By a similar procedure, we also have the following bound.

$$|S_\varepsilon(\hat{\mu}_n, \nu'_J) - S_\varepsilon(\hat{\mu}_n, \nu_J)| \leq 2 \tag{3.12}$$

We recall the bounded difference inequality proved in [McD89]. Let $\mathcal{A}$ be some set. Let $X_1, \ldots, X_n$ be arbitrary independent random variables on set $\mathcal{A}$. Let $\phi : \mathcal{A}^n \to \mathbb{R}$ satisfies the bounded difference assumption as follows.

$$\begin{aligned} &\text{if } \exists c_1, \ldots, c_n \geq 0 \ \text{ s.t. } \forall i, 1 \leq i \leq n \\ &\quad \sup_{x_1,\ldots,x_n,x'_i \in \mathcal{A}} |\phi(x_1, \ldots, x_i, \ldots, x_n) - \phi(x_1, \ldots, x'_i, \ldots, x_n)| \leq c_i \end{aligned} \tag{3.13}$$

Then $\forall t > 0$,

$$\mathbf{P}\left\{|\phi(X_1, \ldots, X_n) - \mathbb{E}[\phi(X_1, \ldots, X_n)]| \geqslant t\right\} \leqslant 2e^{-\frac{2t^2}{\sum_{i=1}^n c_i^2}}. \tag{3.14}$$

By applying the bounded difference inequality, we finally have the result in the proposition.

$$\mathbf{P}\left[\left|\hat{S}_{\varepsilon,n}-\mathbf{E}\left[\hat{S}_{\varepsilon,n}\right]\right|\geq s\right]\leq 2\exp\left(-\frac{2s^2}{n\cdot\left(\frac{2}{n}\right)^2+J\cdot(2^2)}\right)=2\exp\left(-\frac{s^2}{2\left(\frac{1}{n}+J\right)}\right) \tag{3.15}$$

We can also write this bound in the following form.

$$\left|\hat{S}_{\varepsilon,n}-\mathbf{E}\left[\hat{S}_{\varepsilon,n}\right]\right|\lesssim\sqrt{\ln(2/\delta)}\cdot\sqrt{\frac{1}{n}+J} \tag{3.16}$$

with probability $1-\delta$. □

3.1.6 A bound for $\left|\mathbf{E}\left[\hat{S}^{(t)}_{\varepsilon,n}\right]-\hat{S}_{\varepsilon,n}\right|$

We first consider a decomposition of approximation error of $\hat{S}^{(t)}_{\varepsilon,n}$ with respect to $\hat{S}_{\varepsilon,n}$.

$$\begin{aligned}
&\left|\mathbf{E}\left[\hat{S}^{(t)}_{\varepsilon,n}\right]-\hat{S}_{\varepsilon,n}\right| \\
=&\left|\mathbf{E}\left[\bar{H}^1_\varepsilon(\mathbf{v}_1^{(t)})-\frac{1}{2}\left(\bar{H}^2_\varepsilon(\mathbf{v}_2^{(t)})+\bar{H}^3_\varepsilon(\mathbf{v}_3^{(t)})\right)\right]-\left(\bar{H}^1_\varepsilon(\mathbf{v}_1^*)-\frac{1}{2}\left(\bar{H}^2_\varepsilon(\mathbf{v}_2^*)+\bar{H}^3_\varepsilon(\mathbf{v}_3^*)\right)\right)\right| \\
\leq&\left|\mathbf{E}\left[\bar{H}^1_\varepsilon(\mathbf{v}_1^{(t)})\right]-\bar{H}^1_\varepsilon(\mathbf{v}_1^*)\right|+\frac{1}{2}\left|\mathbf{E}\left[\bar{H}^2_\varepsilon(\mathbf{v}_2^{(t)})\right]-\bar{H}^2_\varepsilon(\mathbf{v}_2^*)\right|+\frac{1}{2}\left|\mathbf{E}\left[\bar{H}^3_\varepsilon(\mathbf{v}_3^{(t)})\right]-\bar{H}^3_\varepsilon(\mathbf{v}_3^*)\right|
\end{aligned} \tag{3.17}$$

where

$$\begin{aligned}
W_\varepsilon(\hat{\mu}_n,\nu_J)&=\max_{\mathbf{v}_1\in\mathbb{R}^J}\bar{H}^1_\varepsilon(\mathbf{v}_1)=\frac{1}{n}\sum_{i=1}^n\bar{h}^1_\varepsilon(x_i,\mathbf{v}_1),\\
\bar{h}^1_\varepsilon(x,\mathbf{v})&=\sum_{j=1}^J\mathbf{v}_j\boldsymbol{\nu}_j-\varepsilon\ln\left(\sum_{j=1}^J\exp\left(\frac{\mathbf{v}_j-c\left(x,y_j\right)}{\varepsilon}\right)\boldsymbol{\nu}_j\right)-\varepsilon,\\
W_\varepsilon(\hat{\mu}_n,\hat{\mu}_n)&=\max_{\mathbf{v}_2\in\mathbb{R}^n}\bar{H}^2_\varepsilon(\mathbf{v}_2)=\frac{1}{n}\sum_{i=1}^n\bar{h}^2_\varepsilon(x_i,\mathbf{v}_2),\\
\bar{h}^2_\varepsilon(x,\mathbf{v})&=\frac{1}{n}\sum_{i=1}^n\mathbf{v}_i-\varepsilon\ln\left(\frac{1}{n}\sum_{i=1}^n\exp\left(\frac{\mathbf{v}_i-c\left(x,x_i\right)}{\varepsilon}\right)\right)-\varepsilon,\\
W_\varepsilon(\nu_J,\nu_J)&=\max_{\mathbf{v}_3\in\mathbb{R}^J}\bar{H}^3_\varepsilon(\mathbf{v}_3)=\sum_{j=1}^J\bar{h}^3_\varepsilon(y_j,\mathbf{v}_3)\boldsymbol{\nu}_j,\\
\bar{h}^3_\varepsilon(y,\mathbf{v})&=\sum_{j=1}^J\mathbf{v}_j\boldsymbol{\nu}_j-\varepsilon\ln\left(\sum_{j=1}^J\exp\left(\frac{\mathbf{v}_j-c\left(y,y_j\right)}{\varepsilon}\right)\boldsymbol{\nu}_j\right)-\varepsilon.
\end{aligned} \tag{3.18}$$

Then we study each term separately.

In [SRB13], the author presented the convergence result of SAG algorithm. Assuming that $f_i, i=1...n$ in problem (2.15) are convex and differentiable and their gradients f_i' are Lipchitz-continuous with constant L, which means that

$$\forall x,y\in\mathbb{R}^{\mathsf{d}},\|f_i'(x)-f_i'(y)\|\leq L\|x-y\|. \tag{3.19}$$

We also assume that there exists at least one minimizer x^* to attain the optimal value. Let $\bar{x}^t$ denote the average iterate $\frac{1}{t}\sum_{i=0}^{t-1} x^i$. Let σ^2 denote the variance of the gradient norm at x^*, $\frac{1}{n}\sum_{i=1}^{n}\|f_i'(x^*)\|^2$. Then with the step size $\alpha_t = \frac{1}{16L}$, it holds,

$$\mathbf{E}\left[g\left(\bar{x}^t\right)\right] - g\left(x^*\right) \leqslant \frac{32n}{t} C_0 \tag{3.20}$$

where when y_i^0 are initialized with 0,

$$C_0 = g\left(x^0\right) - g\left(x^*\right) + \frac{4L}{n}\left\|x^0 - x^*\right\|^2 + \frac{\sigma^2}{16L} \tag{3.21}$$

The author also indicate that this bound also holds for x^t instead of $\bar{x}^t$.

In our case, the sum of function to maximize is $\bar{H}_\varepsilon$ in the discrete semi-dual formulation of regularized OT (2.10). We have

$$\bar{H}_\varepsilon(\mathbf{v}) = \sum_{i=1}^{n} \boldsymbol{\mu}_i \bar{h}_\varepsilon(x_i, \mathbf{v}) \tag{3.22}$$

where $\nabla_v \bar{h}_\varepsilon(x_i, \mathbf{v})$ are L-Lipchitz-continuous with L upper bounded by $\max_i \frac{\boldsymbol{\mu}_i}{\varepsilon}$, according to [GCPB16].

Therefore we obtain the bound for $\left|\mathbf{E}\left[\hat{S}_{\varepsilon,n}^{(t)}\right] - \hat{S}_{\varepsilon,n}\right|$, by summing the following bounds.

$$\begin{aligned}
\left|\mathbf{E}\left[\bar{H}_\varepsilon^1(\mathbf{v}_1^{(t)})\right] - \bar{H}_\varepsilon^1(\mathbf{v}_1^*)\right| &\leq \frac{32n}{t}\left(\bar{H}_\varepsilon^1(\mathbf{v}_1^*) - \bar{H}_\varepsilon^1(\mathbf{v}_1^0) + \frac{4}{n^2\varepsilon}\|\mathbf{v}_1^0 - \mathbf{v}_1^*\|^2 + \frac{\sigma_1^2 n\varepsilon}{16}\right) \\
\frac{1}{2}\left|\mathbf{E}\left[\bar{H}_\varepsilon^2(\mathbf{v}_2^{(t)})\right] - \bar{H}_\varepsilon^2(\mathbf{v}_2^*)\right| &\leq \frac{16n}{t}\left(\bar{H}_\varepsilon^2(\mathbf{v}_2^*) - \bar{H}_\varepsilon^2(\mathbf{v}_2^0) + \frac{4}{n^2\varepsilon}\|\mathbf{v}_2^0 - \mathbf{v}_2^*\|^2 + \frac{\sigma_2^2 n\varepsilon}{16}\right) \\
\frac{1}{2}\left|\mathbf{E}\left[\bar{H}_\varepsilon^3(\mathbf{v}_3^{(t)})\right] - \bar{H}_\varepsilon^3(\mathbf{v}_3^*)\right| &\leq \frac{16J}{t}\left(\bar{H}_\varepsilon^3(\mathbf{v}_3^*) - \bar{H}_\varepsilon^3(\mathbf{v}_3^0) + \frac{4}{J\varepsilon}\|\mathbf{v}_3^0 - \mathbf{v}_3^*\|^2 + \frac{\sigma_3^2\varepsilon}{16}\right)
\end{aligned} \tag{3.23}$$

3.1.7 Summing up for fixed ε

By combining the results in section 3.1.3, 3.1.4, 3.1.5 and 3.1.6, we obtain the regret bound of $\hat{S}_{\varepsilon,n}^{(t)}$ with respect to W_2^2..

Proposition 3. *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ be a continuous measure, and $\nu_J \in \mathcal{P}(\mathbb{R}^d)$ be a discrete measure of size J whose weights are random. Both of them are concentrated on a set of diameter 1. Let $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ be a empirical measure of μ where $x_i, i = 1...n$ are points independently sampled from μ. Let $d' = 2\lfloor d/2 \rfloor$. Then*

$$\begin{aligned}
\sum_{t=1}^{T}\left|\mathbf{E}\left[\hat{S}_{\varepsilon,n}^{(t)}\right] - W_2^2\right| \lesssim (\ln(T)+1)\left(C_{11}n^{-1}\varepsilon^{-1} + C_{12}\varepsilon^{-1} + C_{21}n^2\varepsilon + C_{22}\varepsilon + C_{31}n + C_{32}\right) \\
+ T\left(\varepsilon^2 + n^{-1/2} + n^{-1/2}\varepsilon^{-d'/2} + C_4\sqrt{n^{-1}+J}\right)
\end{aligned} \tag{3.24}$$

where

$$
\begin{aligned}
C_{11} &= 128\|\mathbf{v}_1^0 - \mathbf{v}_1^*\|^2 + 64\|\mathbf{v}_2^0 - \mathbf{v}_2^*\|^2, \\
C_{12} &= 64\|\mathbf{v}_3^0 - \mathbf{v}_3^*\|^2, \\
C_{21} &= 2\sigma_1^2 + \sigma_2^2, \\
C_{22} &= J\sigma_3^2, \\
C_{31} &= 32\left(\bar{H}_\varepsilon^1(\mathbf{v}_1^*) - \bar{H}_\varepsilon^1(\mathbf{v}_1^0)\right) + 16\left(\bar{H}_\varepsilon^2(\mathbf{v}_2^*) - \bar{H}_\varepsilon^2(\mathbf{v}_2^0)\right), \\
C_{32} &= 16J\left(\bar{H}_\varepsilon^3(\mathbf{v}_3^*) - \bar{H}_\varepsilon^3(\mathbf{v}_3^0)\right), \\
C_4 &= \sqrt{\ln\left(\frac{2}{\delta}\right)}.
\end{aligned}
\tag{3.25}
$$

Proof. We have the error of $\hat{S}_{\varepsilon,n}^{(t)}$ with respect to W_2^2,

$$
\left|\mathbf{E}\left[\hat{S}_{\varepsilon,n}^{(t)}\right] - W_2^2\right| \leq \left|\mathbf{E}\left[\hat{S}_{\varepsilon,n}^{(t)}\right] - \hat{S}_{\varepsilon,n}\right| + \left|\hat{S}_{\varepsilon,n} - W_2^2\right|. \tag{3.26}
$$

For the first absolute value on the right hand side, its bound can be written in the following form.

$$
\begin{aligned}
\sum_{t=1}^{T}\left|\mathbf{E}\left[\hat{S}_{\varepsilon,n}^{(t)}\right] - \hat{S}_{\varepsilon,n}\right| &\leq \left(\sum_{t=1}^{T}\frac{1}{t}\right)\left(C_{11}n^{-1}\varepsilon^{-1} + C_{12}\varepsilon^{-1} + C_{21}n^2\varepsilon + C_{22}\varepsilon + C_{31}n + C_{32}\right) \\
&< (\ln(T)+1)\left(C_{11}n^{-1}\varepsilon^{-1} + C_{12}\varepsilon^{-1} + C_{21}n^2\varepsilon + C_{22}\varepsilon + C_{31}n + C_{32}\right)
\end{aligned}
\tag{3.27}
$$

where

$$
\begin{aligned}
C_{11} &= 128\|\mathbf{v}_1^0 - \mathbf{v}_1^*\|^2 + 64\|\mathbf{v}_2^0 - \mathbf{v}_2^*\|^2, \\
C_{12} &= 64\|\mathbf{v}_3^0 - \mathbf{v}_3^*\|^2, \\
C_{21} &= 2\sigma_1^2 + \sigma_2^2, \\
C_{22} &= J\sigma_3^2, \\
C_{31} &= 32\left(\bar{H}_\varepsilon^1(\mathbf{v}_1^*) - \bar{H}_\varepsilon^1(\mathbf{v}_1^0)\right) + 16\left(\bar{H}_\varepsilon^2(\mathbf{v}_2^*) - \bar{H}_\varepsilon^2(\mathbf{v}_2^0)\right), \\
C_{32} &= 16J\left(\bar{H}_\varepsilon^3(\mathbf{v}_3^*) - \bar{H}_\varepsilon^3(\mathbf{v}_3^0)\right).
\end{aligned}
\tag{3.28}
$$

For the second absolute value on the right hand side, we can find its bound by using the bounds in subsection 3.1.3, 3.1.4 and 3.1.5.

$$
T\left|\hat{S}_{\varepsilon,n} - W_2^2\right| \lesssim T\left(\varepsilon^2 + +n^{-1/2} + n^{-1/2}\varepsilon^{-d'/2} + C_4\sqrt{n^{-1}+J}\right) \tag{3.29}
$$

with probability $1-\delta$, where

$$
C_4 = \sqrt{\ln\left(\frac{2}{\delta}\right)}. \tag{3.30}
$$

The desired result follows. □

Let $B(n,\varepsilon)$ denote the statistical bound of the regret of $\hat{S}_{\varepsilon,n}^{(t)}$, which can be seen as a function of n and ε. We want to acquire a pair of n and ε in function of T that minimizes B. However it is difficult to find an explicit expression. To simplify the problem, we let n and ε be polynomials of T and then replace them back into the

expression of B. The aim is to make the order of T in the expression of B within our acceptable range.

Let $\varepsilon = A_\varepsilon T^a$ and $n = A_n T^b$, where $A_\varepsilon, A_n > 0$ and $a, b \in \mathbb{R}$. We rewrite the bound B as follows.

$$\begin{aligned} B_T = (\ln T + 1) \left(\frac{C_{11}}{A_\varepsilon A_n} T^{-(a+b)} + \frac{C_{12}}{A_\varepsilon} T^{-a} + C_{21} A_\varepsilon A_n^2 T^{a+2b} + C_{22} A_\varepsilon T^a + C_{31} A_n T^b + C_{32} \right) \\ + T \left(A_\varepsilon^2 T^{2a} + A_n^{-\frac{1}{2}} A_\varepsilon^{-\frac{d'}{2}} T^{-\frac{d'}{2}a - \frac{1}{2}b} + A_n^{-\frac{1}{2}} T^{-\frac{1}{2}b} + C_4 \sqrt{A_n^{-1} T^{-b} + J} \right) \end{aligned} \tag{3.31}$$

Based on the expression of B_T, we set the following objective.

$$\underset{a,b \in \mathbb{R}}{\text{minimize}} \ \max\{-a-b, -a, a+2b, a, b, 2a+1, -\frac{d'}{2}a - \frac{1}{2}b + 1, -\frac{1}{2}b + 1\} \tag{3.32}$$

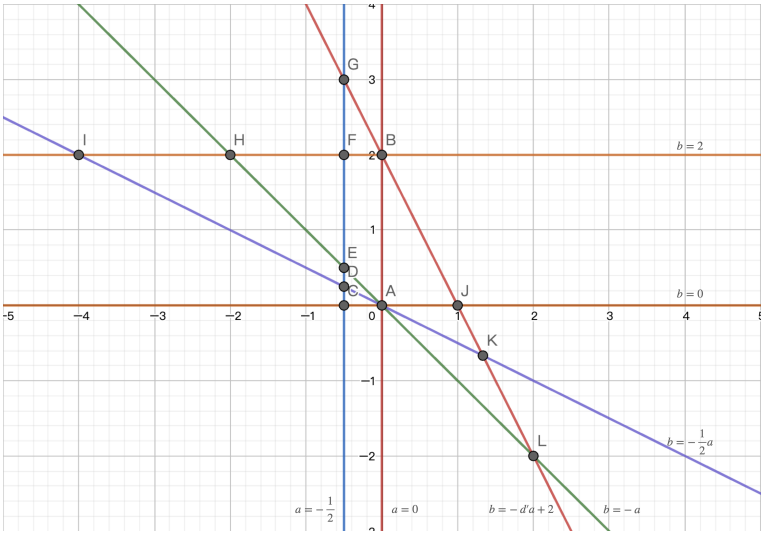


Figure 3.1: Illustration of the objective on $\mathbb{R}^2$

We then investigated each of the points of intersection in the graph, and the results are shown in the table below.

point	a	b	$-a-b$	$-a$	$a+2b$	$2a+1$	$-\frac{1}{2}b-\frac{d'}{2}a+1$	$-\frac{1}{2}b+1$	max
A	0	0	0	0	0	1	1	1	1
B	0	2	-2	0	4	1	0	0	4
C	$-\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{2}$	0	$\frac{d'}{4}+1$	1	$\geq\frac{3}{2}$
D	$-\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{2}$	0	0	$\frac{d'}{4}+\frac{7}{8}$	$\frac{7}{8}$	$\geq\frac{11}{8}$
E	$-\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{d'}{4}+\frac{3}{4}$	$\frac{3}{4}$	$\geq\frac{5}{4}$
F	$-\frac{1}{2}$	2	$-\frac{3}{2}$	$\frac{1}{2}$	$\frac{7}{2}$	0	$\frac{d'}{4}$	0	$\geq\frac{7}{2}$
G	$-\frac{1}{2}$	$\frac{d'}{2}+2$	$-\frac{d'}{2}-\frac{3}{2}$	$\frac{1}{2}$	$d'+\frac{7}{2}$	0	0	$-\frac{d'}{4}$	$\geq\frac{11}{2}$
H	-2	2	0	2	2	-3	d'	0	≥ 2
I	-4	2	2	4	0	-7	$2d'$	0	≥ 4
J	$\frac{2}{d'}$	0	$-\frac{2}{d'}$	$-\frac{2}{d'}$	$\frac{2}{d'}$	$\frac{4}{d'}+1$	0	1	3
K	$\frac{4}{2d'-1}$	$\frac{-2}{2d'-1}$	$\frac{-2}{2d'-1}$	$\frac{-4}{2d'-1}$	0	$\frac{8}{2d'-1}+1$	0	$\frac{1}{2d'-1}+1$	$\frac{11}{3}$
L	$\frac{2}{d'-1}$	$\frac{-2}{d'-1}$	0	$\frac{-2}{d'-1}$	$\frac{-2}{d'-1}$	$\frac{4}{d'-1}+1$	0	$\frac{1}{d'-1}+1$	5

Table 3.1: Analysis of the order of T

As we can see from this Table (3.1), under the previous simplifying assumptions on n and ε, minimizing the order of T in B_T is indeed a difficult problem to balance. The optimal solution is what the first row shows, where the expression of B_T can have a order of $O(T)$ when $a = b = 0$. If we continue this analysis, we can make more complex assumptions about ε and n by introducing more variables, such as $\varepsilon = A_{\varepsilon,1}T^{a_1} + A_{\varepsilon,2}T^{a_2}$ and $n = A_{n,1}T^{b_1} + A_{n,2}T^{b_2}$. Since the process will be more complicated, we will not expand it here.

3.1.8 Summing up for decreasing ε

Similar to the case of fixed ε, we combine the results in section 3.1.3, 3.1.4, 3.1.5 and 3.1.6, but it is important to note that here we replace ε by ε_t, which is a decreasing function of t.

Proposition 4. *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ be a continuous measure, and $\nu_J \in \mathcal{P}(\mathbb{R}^d)$ be a discrete measure of size J whose weights are random. Both of them are concentrated on a set of diameter 1. Let $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{x_i}$ be a empirical measure of μ where $x_i, i = 1...n$ are points independently sampled from μ. Let $d' = 2\lfloor d/2 \rfloor$. Let $\varepsilon_t = \frac{\varepsilon_0}{\sqrt{t}}$.*

Then

$$\begin{aligned}\sum_{t=1}^{T}\left|\hat{S}^{(t)}_{\varepsilon_t,n}-W_2^2\right| \lesssim & (\ln T+1)(C_{31}n+C_{32}+\varepsilon_0^2)\\ &+\left((T+1)^{\frac{d'}{4}+1}-1\right)n^{-\frac{1}{2}}\varepsilon_0^{-\frac{d'}{2}}\cdot\frac{4}{d'+4}\\ &+T\left(n^{-\frac{1}{2}}+C_4\sqrt{n^{-1}+J}\right)\\ &+\left(2T^{\frac{1}{2}}-1\right)(C_{11}n^{-1}+C_{12})\varepsilon_0^{-1}\\ &+\left(3-2T^{-\frac{1}{2}}\right)(C_{21}n^2+C_{22})\varepsilon_0\end{aligned} \tag{3.33}$$

where

$$\begin{aligned}C_{11}&=128\|\mathbf{v}_1^0-\mathbf{v}_1^*\|^2+64\|\mathbf{v}_2^0-\mathbf{v}_2^*\|^2,\\ C_{12}&=64\|\mathbf{v}_3^0-\mathbf{v}_3^*\|^2,\\ C_{21}&=2\sigma_1^2+\sigma_2^2,\\ C_{22}&=J\sigma_3^2,\\ C_{31}&=32\left(\bar{H}^1_\varepsilon(\mathbf{v}_1^*)-\bar{H}^1_\varepsilon(\mathbf{v}_1^0)\right)+16\left(\bar{H}^2_\varepsilon(\mathbf{v}_2^*)-\bar{H}^2_\varepsilon(\mathbf{v}_2^0)\right),\\ C_{32}&=16J\left(\bar{H}^3_\varepsilon(\mathbf{v}_3^*)-\bar{H}^3_\varepsilon(\mathbf{v}_3^0)\right),\\ C_4&=\sqrt{\ln\left(\frac{2}{\delta}\right)}.\end{aligned} \tag{3.34}$$

Proof. First of all, we decompose $\left|\hat{S}^{(t)}_{\varepsilon_t,n}-W_2^2\right|$ as in the previous subsection.

$$\left|\mathbf{E}\left[\hat{S}^{(t)}_{\varepsilon_t,n}\right]-W_2^2\right|\le\left|\mathbf{E}\left[\hat{S}^{(t)}_{\varepsilon_t,n}\right]-\hat{S}_{\varepsilon_t,n}\right|+\left|\hat{S}_{\varepsilon_t,n}-\mathbf{E}\left[\hat{S}_{\varepsilon_t,n}\right]\right|+\mathbf{E}\left[\left|\hat{S}_{\varepsilon_t,n}-S_{\varepsilon_t}\right|\right]+\left|S_{\varepsilon_t}-W_2^2\right|. \tag{3.35}$$

Then we consider them sum along the iteration t separately.

$$\sum_{t=1}^{T}\left|S_{\varepsilon_t}-W_2^2\right|\lesssim\sum_{t=1}^{T}\varepsilon_t^2=\sum_{t=1}^{T}\frac{\varepsilon_0^2}{t}<\varepsilon_0^2(\ln T+1), \tag{3.36}$$

$$\begin{aligned}\sum_{t=1}^{T}\mathbf{E}\left[\left|\hat{S}_{\varepsilon_t,n}-S_{\varepsilon_t}\right|\right]&\lesssim Tn^{-\frac{1}{2}}+n^{-\frac{1}{2}}\sum_{t=1}^{T}\varepsilon_t^{-\frac{d'}{2}}=Tn^{-\frac{1}{2}}+n^{-\frac{1}{2}}\varepsilon_0^{-\frac{d'}{2}}\sum_{t=1}^{T}t^{\frac{d'}{4}}\\ &<Tn^{-\frac{1}{2}}+n^{-\frac{1}{2}}\varepsilon_0^{-\frac{d'}{2}}\cdot\frac{4}{d'+4}\left((T+1)^{\frac{d'}{4}+1}-1\right),\end{aligned} \tag{3.37}$$

$$\sum_{t=1}^{T}\left|\hat{S}_{\varepsilon_t,n}-\mathbf{E}\left[\hat{S}_{\varepsilon_t,n}\right]\right|\lesssim C_4T\sqrt{\frac{1}{n}+J}, \tag{3.38}$$

with probability $1-\delta$. We deduce that

$$\begin{aligned}\sum_{t=1}^{T}\left|\mathbf{E}\left[\hat{S}^{(t)}_{\varepsilon_t,n}\right]-\hat{S}_{\varepsilon_t,n}\right| \leq & (C_{11}n^{-1}+C_{12})\sum_{t=1}^{T}\left(\frac{1}{t\varepsilon_t}\right)+(C_{21}n^2+C_{22})\sum_{t=1}^{T}\left(\frac{\varepsilon_t}{t}\right)\\ &+(C_{31}n+C_{32})\sum_{t=1}^{T}\frac{1}{t}\\ =&(C_{11}n^{-1}+C_{12})\varepsilon_0^{-1}\sum_{t=1}^{T}\left(t^{-\frac{1}{2}}\right)+(C_{21}n^2+C_{22})\varepsilon_0\sum_{t=1}^{T}\left(t^{-\frac{3}{2}}\right)\\ &+(C_{31}n+C_{32})\sum_{t=1}^{T}t^{-1}\\ <&(C_{11}n^{-1}+C_{12})\varepsilon_0^{-1}\left(2T^{\frac{1}{2}}-1\right)+(C_{21}n^2+C_{22})\varepsilon_0\left(3-2T^{-\frac{1}{2}}\right)\\ &+(C_{31}n+C_{32})(\ln(T)+1).\end{aligned} \tag{3.39}$$

Combining these inequalities provide a proof of the desired result. □

Similar to the previous setting, we also want to get an expression for n in function of T which is difficult to compute. Therefore we perform the same simplification as in section 3.1.7. Let $n = AT^a$ where $A > 0$ and $a \in \mathbb{R}$. Let $B(n)$ denote the statistical bound of regret of $\hat{S}^{(t)}_{\varepsilon_t,n}$. We rewrite the bound B_T as follows.

$$\begin{aligned}B_T =&(\ln T+1)(C_{31}AT^a+C_{32}+\varepsilon_0^2)\\ &+\left((T+1)^{\frac{d'}{4}+1}-1\right)A^{-\frac{1}{2}}T^{-\frac{a}{2}}\varepsilon_0^{-\frac{d'}{2}}\cdot\frac{4}{d'+4}\\ &+T\left(A^{-\frac{1}{2}}T^{-\frac{a}{2}}+C_4\sqrt{A^{-1}T^{-a}+J}\right)\\ &+\left(2T^{\frac{1}{2}}-1\right)(C_{11}A^{-1}T^{-a}+C_{12})\varepsilon_0^{-1}\\ &+\left(3-2T^{-\frac{1}{2}}\right)(C_{21}A^2T^{2a}+C_{22})\varepsilon_0\end{aligned} \tag{3.40}$$

Based on the expression of B_T, we set the following objective.

$$\underset{a\in\mathbb{R}}{\text{minimize}}\ \max\left\{a,\frac{d'}{4}+1-\frac{a}{2},\frac{1}{2}-a,2a\right\} \tag{3.41}$$

a	$\frac{d'}{4}+1-\frac{a}{2}$	$\frac{1}{2}-a$	$2a$	max
0	$\frac{d'}{4}+1$	$\frac{1}{2}$	0	$\geq\frac{3}{2}$
$\frac{1}{4}$	$\frac{d'}{4}+\frac{7}{8}$	$\frac{1}{4}$	$\frac{1}{2}$	$\geq\frac{11}{8}$
$\frac{d'}{6}+\frac{2}{3}$	$\frac{d'}{6}+\frac{2}{3}$	$-\frac{d'}{6}-\frac{1}{6}$	$\frac{d'}{3}+\frac{4}{3}$	≥ 2
$-\frac{d'}{2}-1$	$\frac{d'}{2}+\frac{3}{2}$	$\frac{d'}{2}+\frac{3}{2}$	$-d'-2$	$\geq\frac{5}{2}$
$\frac{d'}{10}+\frac{2}{5}$	$\frac{d'}{5}+\frac{4}{5}$	$-\frac{d'}{10}+\frac{1}{10}$	$\frac{d'}{5}+\frac{4}{5}$	$\geq\frac{6}{5}$
$\frac{1}{6}$	$\frac{d}{4}+\frac{11}{12}$	$\frac{1}{3}$	$\frac{1}{3}$	$\geq\frac{17}{12}$

Table 3.2: Analysis of order of T

As can be seen from this table, in each case there is an order that increases with d', so as long as d' is particularly large, it is not possible to find a solution for a ensuring that the order of this bound is not too large. This result may be due to our overly simple assumption about n, which can be further ameliorated by making a more complicated assumption on n.

3.2 Regret bound for Regularized Wasserstein distance with fixed ε

3.2.1 Notations

The notations used in this subsection are listing here:

- μ: a continuous measure in $\mathcal{P}(\mathbb{R}^d)$.
- $\nu_J = \sum_{j=1}^{J} \boldsymbol{\nu}_j \delta_{y_j}$: a discrete measure in $\mathcal{P}(\mathbb{R}^d)$ with sample size J, where $\boldsymbol{\nu} \in \Sigma_J$ is a weight vector for ν_J and the weights are random.
- W_2^2: $W_2^2(\mu, \nu_J)$, squared Wasserstein distance between μ and ν_J.
- W_ε: $W_\varepsilon(\mu, \nu_J)$, regularized squared Wasserstein distance between μ and ν_J.
- $W_\varepsilon^{(t)}$: estimation of W_ε at iteration t, using ASGD algorithm (2).

3.2.2 Introduction

In this subsection, we investigate a bound for $\sum_{t=1}^{T} \left| \mathbf{E}\left[\hat{W}_\varepsilon^{(t)}\right] - W_2^2 \right|$, where $\hat{W}_\varepsilon^{(t)}$ is an approximation of regularized Wasserstein distance using ASGD algorithm (2) at iteration t.

3.2.3 A bound for $\left|W_\varepsilon - W_2^2\right|$

In [CRL+20], the author stated in the first section that $W_\varepsilon(\mu, \nu)$ approximates $W_2^2(\mu, \nu)$ with an error of $O(\varepsilon \ln(1/\varepsilon))$. This error can be written as follows.

$$\left|W_\varepsilon(\mu, \nu) - W_2^2(\mu, \nu)\right| \leq O\left(\varepsilon \ln\left(\frac{1}{\varepsilon}\right)\right). \tag{3.42}$$

3.2.4 A bound for $\left|\mathbf{E}\left[W_\varepsilon^{(t)} - W_\varepsilon\right]\right|$

Here we consider the semi-dual formulation of the regularized Wasserstein distance in the semi-discrete case (2.13) as the objective function and we apply the ASGD algorithm (2) to find an solution $\mathbf{v}^{(t)} \in \mathbb{R}^J$ after t iteration, which is a simple average

of iterates of SGD. Let $\mathbf{v}^* \in \mathbb{R}^J$ denote the optimal solution. We have the relation between the notations and the formulation as follows.

$$\begin{aligned} \mathbf{E}\left[W_\varepsilon^{(t)}\right] &= \mathbf{E}\left[\bar{h}_\varepsilon(X, \mathbf{v}^{(t)})\right], \\ W_\varepsilon &= \mathbf{E}\left[\bar{h}_\varepsilon(X, \mathbf{v}^*)\right] \end{aligned} \tag{3.43}$$

where the expression of $\bar{h}_\varepsilon$ is defined in (2.11). Then we derive a bound of optimization error of Algorithm (2) by adapting Theorem 2 in [SZ12].

Proposition 5. *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ be a continuous measure, and $\nu_J \in \mathcal{P}(\mathbb{R}^d)$ be a discrete measure of size J whose weights are random. Both of them are concentrated on a set of diameter 1. For some constants D, it holds that $\sup_{\mathbf{v},\mathbf{v}' \in \mathbb{R}^J} \|\mathbf{v} - \mathbf{v}'\| \leq D$. With step sizes $\alpha_k = \frac{\alpha}{\sqrt{t}}$ where $\alpha > 0$ is a constant. Then for any $t > 0$, it holds that*

$$\left|\mathbf{E}\left[W_\varepsilon^{(t)} - W_\varepsilon\right]\right| \leq 2\left(\frac{D^2}{\alpha} + 4\alpha\right)\frac{2 + \ln t}{\sqrt{t}} \tag{3.44}$$

where $W_\varepsilon^{(t)}$ is accessed by simple averaging iterates at t.

Proof. In [SZ12], Theorem 2, the authors showed a bound of expected error of a general convex function F in a individual iterate of SGD $\tilde{\mathbf{v}}$ with a step size of $\alpha_k = \frac{\alpha}{\sqrt{k}}$, where $\alpha > 0$.

$$\mathbf{E}\left[F(\tilde{\mathbf{v}}^{(t)}) - F(\mathbf{v}^*)\right] \leq \left(\frac{D^2}{\alpha} + \alpha G^2\right)\frac{2 + \ln t}{\sqrt{t}} \tag{3.45}$$

where D, G are some constants that hold $\sup_{\mathbf{v},\mathbf{v}' \in \mathbb{R}^J} \|\mathbf{v} - \mathbf{v}'\| \leq D$ and $\mathbf{E}\left[\|\hat{\mathbf{g}}_t\|^2\right] \leq G^2$ and $\hat{\mathbf{g}}_t$ is a vector at iteration t whose expectation is subgradient of F.

We have the relation between the output of ASGD (2) and the individual iterate of SGD as follows.

$$\mathbf{v}^{(t)} = \frac{1}{t}\sum_{\ell=1}^{t} \tilde{\mathbf{v}}^{(\ell)} \tag{3.46}$$

Then we can have a bound of expected error of $F(\mathbf{v}^{(t)})$ as follows.

$$\begin{aligned}
\mathbf{E}\left[F(\mathbf{v}^{(t)}) - F(\mathbf{v}^*)\right] &= \mathbf{E}\left[F\left(\frac{1}{t}\sum_{\ell=1}^{t}\tilde{\mathbf{v}}^{(\ell)}\right) - F(\mathbf{v}^*)\right] \\
&\leq \mathbf{E}\left[\frac{1}{t}\sum_{\ell=1}^{t}F\left(\tilde{\mathbf{v}}^{(\ell)}\right) - F(\mathbf{v}^*)\right] \\
&\leq \frac{1}{t}\sum_{\ell=1}^{t}\mathbf{E}\left[F\left(\tilde{\mathbf{v}}^{(\ell)}\right) - F(\mathbf{v}^*)\right] \\
&= \frac{1}{t}\sum_{\ell=1}^{t}\left(\frac{D^2}{\alpha} + \alpha G^2\right)\frac{2+\ln\ell}{\sqrt{\ell}} \\
&\leq \left(\frac{D^2}{\alpha} + \alpha G^2\right)\frac{2+\ln t}{t}\sum_{\ell=1}^{t}\ell^{-\frac{1}{2}} \\
&\leq \left(\frac{D^2}{\alpha} + \alpha G^2\right)\frac{2+\ln t}{t}(2t^{\frac{1}{2}}) \\
&= 2\left(\frac{D^2}{\alpha} + \alpha G^2\right)\frac{2+\ln t}{\sqrt{t}}
\end{aligned} \tag{3.47}$$

In our case, $\hat{\mathbf{g}}_t$ is actually $\nabla_v \bar{h}_\varepsilon(x, \mathbf{v})$ (2.12) which can be seen as the difference between two simplex of size J. Therefore we acquire a value of the bound $G^2 = 4$. □

3.2.5 Summing up

Finally we combine the results of section 3.2.3 and 3.2.4 to yield a regret bound of $\hat{W}_\varepsilon^{(t)}$ with respect to W_2^2.

Proposition 6. *Let $\mu \in \mathcal{P}(\mathbb{R}^d)$ be a continuous measure, and $\nu_J \in \mathcal{P}(\mathbb{R}^d)$ be a discrete measure of size J whose weights are random. Both of them are concentrated on a set of diameter 1. For some constants D, it holds that $\sup_{\mathbf{v},\mathbf{v}'\in\mathbb{R}^J}\|\mathbf{v}-\mathbf{v}'\| \leq D$. With step sizes $\alpha_k = \frac{\alpha}{\sqrt{t}}$ where $\alpha > 0$ is a constant. Then*

$$\sum_{t=1}^{T}\left|\mathbf{E}\left[\hat{W}_\varepsilon^{(t)}\right] - W_2^2\right| \lesssim T\varepsilon\ln\left(\frac{1}{\varepsilon}\right) + C_5(2+\ln T)\cdot T^{\frac{1}{2}} \tag{3.48}$$

where

$$C_5 = 4\left(\frac{D^2}{\alpha} + 4\alpha\right) \tag{3.49}$$

Proof. We start by decomposing the error of $\mathbf{E}\left[\hat{W}_\varepsilon^{(t)}\right]$ with respect to W_2^2,

$$\left|\mathbf{E}\left[\hat{W}_\varepsilon^{(t)}\right] - W_2^2\right| \leq \left|\mathbf{E}\left[\hat{W}_\varepsilon^{(t)}\right] - W_\varepsilon\right| + \left|W_\varepsilon - W_2^2\right|. \tag{3.50}$$

Then we consider the sum of bound along the iteration t of each term on the right hand side.

$$\sum_{t=1}^{T}\left|\mathbf{E}\left[\hat{W}_\varepsilon^{(t)}\right] - W_\varepsilon\right| \lesssim T\varepsilon\ln\left(\frac{1}{\varepsilon}\right). \tag{3.51}$$

Let $C_5 = 4\left(\frac{D^2}{\alpha} + 4\alpha\right)$, we have

$$\begin{aligned}\sum_{t=1}^{T} \left|W_\varepsilon - W_2^2\right| &\leq \frac{C_5}{2} \sum_{t=1}^{T} \frac{2 + \ln t}{\sqrt{t}} \\ &\leq C_5(2 + \ln T) \cdot T^{\frac{1}{2}}.\end{aligned} \tag{3.52}$$

Combining these inequalities provide a proof of the desired result. □

As long as we set the hyperparameter $\varepsilon = O(T^a)$ where $a < -\frac{1}{2}$ in advance, we can get a bound of order $O(T^{\frac{1}{2}} \ln T)$. However, after acquiring a solution $\mathbf{v}$ by the ASGD algorithm (2), we still we need to pay the cost of discretization of μ in order access $W_\varepsilon^{(t)}$ since it is an expectation on μ. This cost can be ignored whenever discretization number N is large enough.

Chapter 4

Numerical experiments

In this section, we introduce the approximation of the squared Wasserstein distance $W_2^2(\mu, \nu)$ using the estimator $\hat{S}_\varepsilon$ based on Sinkhorn Divergence under discrete and semi-discrete settings. Each experiment was conducted 100 times and a different random seed was set before each start.

4.1 Calculation of the true value of the squared Wasserstein distance

In order to present the result of our method of approximation and calculate the regret of the estimator $\hat{S}_\varepsilon$, it is necessary to first obtain the real value of the squared Wasserstein distance W_2^2 as a reference.

In the case of the discrete setting, since both I and J are relatively small, Linear Programming can be used to obtain the true value of W_2^2.

However, in the case of semi-discrete setting, since μ is continuously distributed, Linear Programming can no longer be used to calculate the true value of the squared Wasserstein distance. We first tried to obtain a discretized distribution $\hat{\mu}_n$ from μ by sampling n points, then used the SAG algorithm (1) when $\varepsilon = 0$ to compute a maximizer $\mathbf{v}$, and then used it to compute $W_2^2(\hat{\mu}_n, \nu)$ by utilizing (2.10) and (2.11). However, the results of this method are very unstable.

We then used the averaged SGD algorithm (2) when $\varepsilon = 0$ to get a maximizer $\mathbf{v}$. Then we sampled N points from μ to compute a value W close to $W_2^2(\mu, \nu)$ by using the expressions in (2.10) and (2.11). The calculation process is shown in the Algorithm (3). In this experiment, we took $N = 100000$, $J = 10$. The calculation process is shown in the Algorithm (3). The output value of W is much more stable than the previous method.

To verify that the value of W obtained by this method is really close to the true value of the squared Wasserstein distance, we compared the W of samples from two Gaussian distribution and its true value, since we have the explicit formula (2.3) of $W_2^2(\mu, \nu)$ when μ and ν are both Gaussian distribution.

The result are shown in Figure 4.1 and 4.2.

Algorithm 3 Approximation of W_2^2 with large number of samples in semi-discrete setting

Input: $\mu, \nu = \sum_{j=1}^{J} \boldsymbol{\nu}_j \delta_{y_j}, \alpha, N$
Output: W
  $\tilde{\mathbf{v}} \leftarrow \mathbb{0}_J$, $\mathbf{v} \leftarrow \tilde{\mathbf{v}}$
  for $k = 1, 2, ...$ **do**
    Sample x_k from μ
    $\tilde{\mathbf{v}} \leftarrow \tilde{\mathbf{v}} + \frac{\alpha}{\sqrt{k}} \nabla_v \bar{h}_\varepsilon(x_k, \tilde{\mathbf{v}})$
    $\mathbf{v} \leftarrow \frac{1}{k}\tilde{\mathbf{v}} + \frac{k-1}{k}\mathbf{v}$
  end for
  Sample N supports $x_1, ..., x_N$ from μ
  for $i = 1, 2, ..., N$ **do**
    $h_i \leftarrow \sum_{j=1}^{J} \mathbf{v}_j \boldsymbol{\nu}_j + \min_j \left(\|x_i - y_j\|^2 - \mathbf{v}_j \right)$
  end for
  $W \leftarrow \frac{\sum_{i=1}^{N} h_i}{N}$

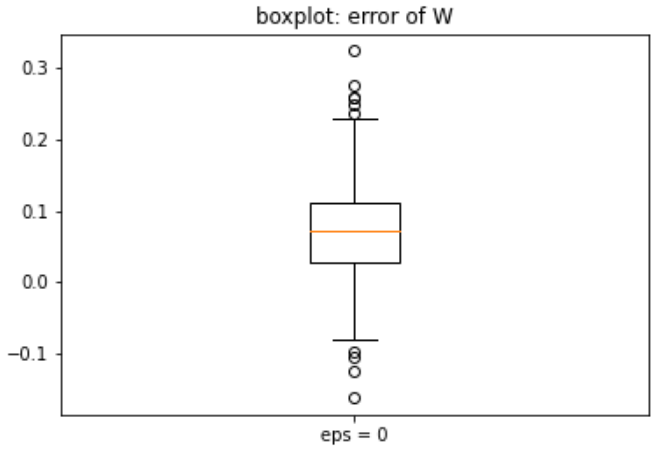


Figure 4.1: Boxplot of error of W

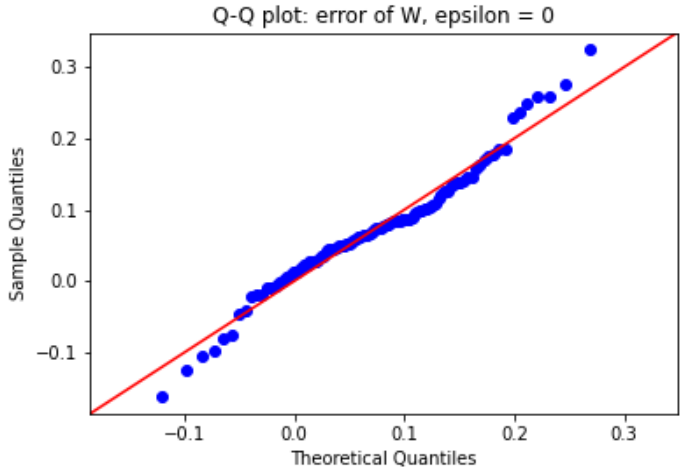


Figure 4.2: Q-Q plot of error of W

4.2 Approximation in discrete case

4.2.1 Distribution used in the experiment and its generation

In the discrete setting, both μ and ν are discrete distributions, which can be express as $\sum_{i=1}^{I} \boldsymbol{\mu}_i \delta_{x_i}$, $\sum_{j=1}^{J} \boldsymbol{\nu}_j \delta_{y_j}$ respectively.

In this experiment, the size of the samples I and J are random integer between 2 and 20, while μ and ν are empirical measure of some samples from two Gaussian mixture ρ_μ and ρ_ν whose parameters are also randomly generated. Then the corresponding supports $x_i \in \mathbb{R}^3, i = 1, ..., I$ and $y_j \in \mathbb{R}^3, j = 1, ..., J$ are sampled from ρ_μ and ρ_ν. Finally the weight vectors $\boldsymbol{\mu} \in \Sigma_I, \boldsymbol{\nu} \in \Sigma_J$ are randomly generated.

4.2.2 Calculation of $\hat{S}_\varepsilon(\mu, \nu)$ with constant ε

In order to calculate $\hat{S}_\varepsilon(\mu, \nu)$, we first need to obtain three maximizers $\mathbf{v}_{\mu,\nu}$, $\mathbf{v}_{\mu,\mu}$ and $\mathbf{v}_{\nu,\nu}$ by solving three transport problems in semi-dual formulation, between (μ, ν), (μ, μ) and (ν, ν) respectively. In this experiment, these maximizers are obtained using the SAG algorithm (1) and ASGD algorithm (2). To balance the speed of convergence with the range of results, I used the hyperparameters presented in table 4.1.

ε	α (SAG)	number of iteration (SAG)	α (ASGD)	number of iteration (ASGD)
0	0.001	50000	0.1	200000
0.001	0.002		0.1	
0.01	0.01		0.2	
0.1	0.1		0.5	

Table 4.1: Details of hyperparameters choosing in discrete case, constant ε. In fact, after several attempts, we found that ASGD requires more iterations to reach convergence compared to SAG. So here we choose a larger number of iteration for ASGD.

Then $W_\epsilon(\hat{\mu}, \hat{\nu})$, $W_\epsilon(\hat{\mu}, \hat{\mu})$ and $W_\epsilon(\hat{\nu}, \hat{\nu})$ are calculated according to the equation (2.10). Since these calculations involve computing $\bar{h}_\varepsilon(x, \mathbf{v})$ for each x_i, it contains a loop of length I, which is very time consuming. Therefore in this experiment, $W_\epsilon(\cdot, \cdot)$ are computed every 100 iterations. Finally, according to equation (2.14), the value of $\hat{S}_\varepsilon(\mu, \nu)$ is also obtained at each 100 iterations. The results of SAG are shown in Figure 4.3, 4.4 and 4.5, while the results of ASGD are shown in Figure 4.6, 4.7 and 4.8. We note that ASGD does not perform significantly better than SAG, but it requires more iteration, hence we only use SAG algorithm in subsequent experiments.

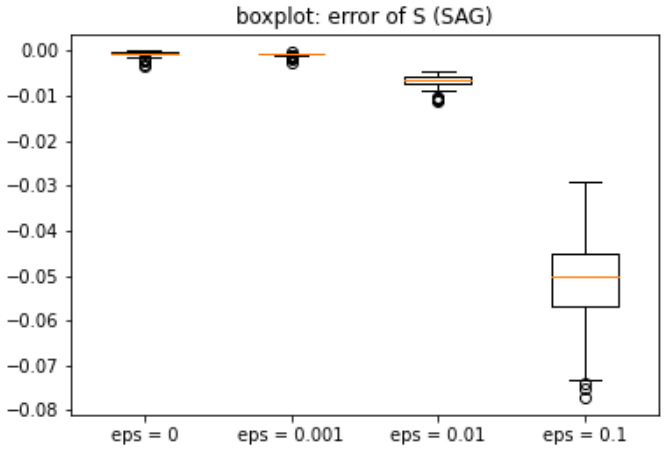
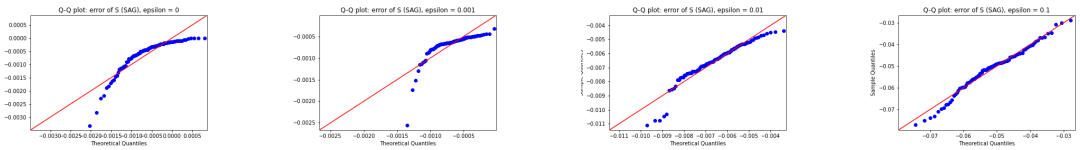
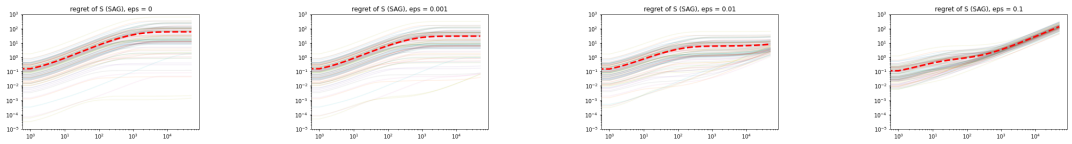


Figure 4.3: Boxplot of error of $\hat{S}_\varepsilon$ (SAG)



(a) $\varepsilon = 0$ (b) $\varepsilon = 0.001$ (c) $\varepsilon = 0.01$ (d) $\varepsilon = 0.1$

Figure 4.4: qqplot of error of $\hat{S}_\varepsilon$ (SAG)



(a) $\varepsilon = 0$ (b) $\varepsilon = 0.001$ (c) $\varepsilon = 0.01$ (d) $\varepsilon = 0.1$

Figure 4.5: Regret of $\hat{S}_\varepsilon$ (SAG)

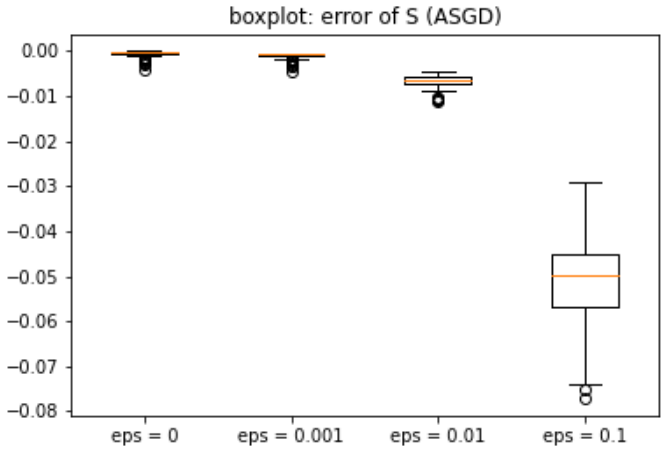
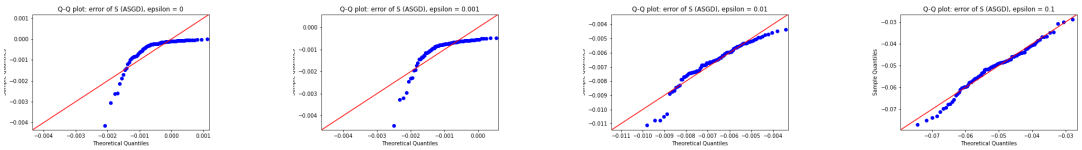
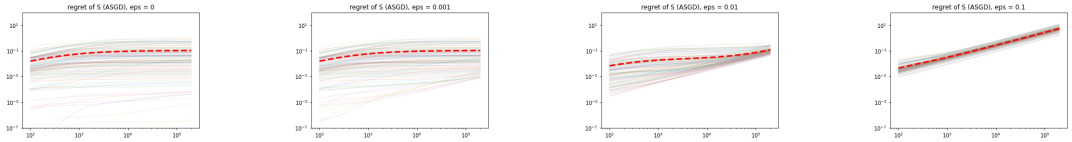


Figure 4.6: Boxplot of error of $\hat{S}_\varepsilon$ (ASGD)



(a) $\varepsilon = 0$ (b) $\varepsilon = 0.001$ (c) $\varepsilon = 0.01$ (d) $\varepsilon = 0.1$

Figure 4.7: qqplot of error of $\hat{S}_\varepsilon$ (ASGD)



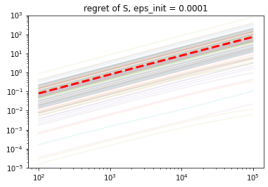
(a) $\varepsilon = 0$ (b) $\varepsilon = 0.001$ (c) $\varepsilon = 0.01$ (d) $\varepsilon = 0.1$

Figure 4.8: Regret of $\hat{S}_\varepsilon$ (ASGD)

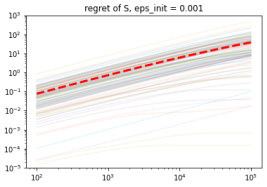
4.2.3 Calculation of $\hat{S}_{\varepsilon_t}(\mu, \nu)$ with decreasing ε_t

The calculation process in the decreasing ε_t case is similar to the process described in section 4.3.2. The only difference is that instead of using a constant ε, here we

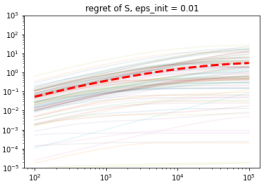
fixed an initial ε_0 and ε_t is updated by $\frac{\varepsilon_0}{\sqrt{t}}$ before each iteration starts. We used a wide range of ε_0, from 10^{-4} to 10^4. Regarding the choice of the learning rate α_t, for simplicity, we chose it to be equal to the epsilon in each iteration. The results are shown in 4.9.



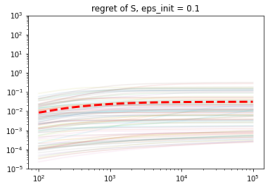
(a) $\varepsilon_0 = 0.0001$



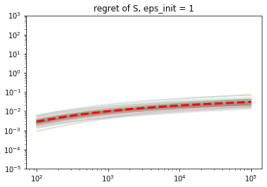
(b) $\varepsilon_0 = 0.001$



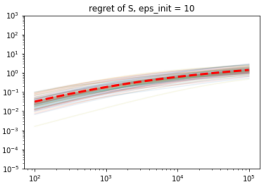
(c) $\varepsilon_0 = 0.01$



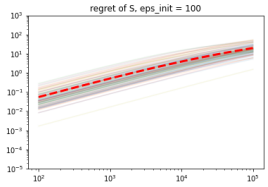
(d) $\varepsilon_0 = 0.1$



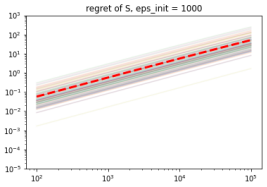
(e) $\varepsilon_0 = 1$



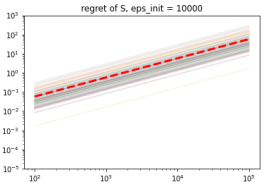
(f) $\varepsilon_0 = 10$



(g) $\varepsilon_0 = 100$



(h) $\varepsilon_0 = 1000$



(i) $\varepsilon_0 = 10000$

Figure 4.9: Regret of $\hat{S}_{\varepsilon_t}$, discrete case

4.3 Approximation in semi-discrete case

4.3.1 Distribution used in the experiment and its generation

In the semi-discrete setting, μ is a continuous distribution, while ν is a discrete distribution. In this experiment, μ is a Gaussian mixture whose parameters are randomly generated and its empirical measure is $\hat{\mu}_n = \sum_{i=1}^{n} \mu_i \delta_{x_i}$ where $n = 1000$ is the size of sample. While $\nu = \sum_{j=1}^{J} \boldsymbol{\nu}_j \delta_{y_j}$ is an empirical measure whose samples are from another Gaussian mixture. Its sample size J is a random integer between 2 and 20. The weight vector $\boldsymbol{\nu} \in \Sigma_J$ are also randomly generated.

4.3.2 Calculation of $\hat{S}_\varepsilon(\mu, \nu)$ with constant ε

The procedure of calculation of $\hat{S}_\varepsilon(\mu, \nu)$ is similar to the one described in section . However, instead of using μ, we use $\hat{\mu}_n$ in the calculation. The hyperparameters are shown in the table 4.2. The results are shown in 4.10, 4.11 and 4.12.

ε	α for $\mathbf{v}_{\hat{\mu}_n,\nu}$	α for $\mathbf{v}_{\hat{\mu}_n,\hat{\mu}_n}$	α for $\mathbf{v}_{\nu,\nu}$	number of iteration
0.001	0.0005	0.001	0.05	50000
0.01		0.005		
0.1		0.05		

Table 4.2: Details of hyperparameters choosing in semi-discrete case, constant ε

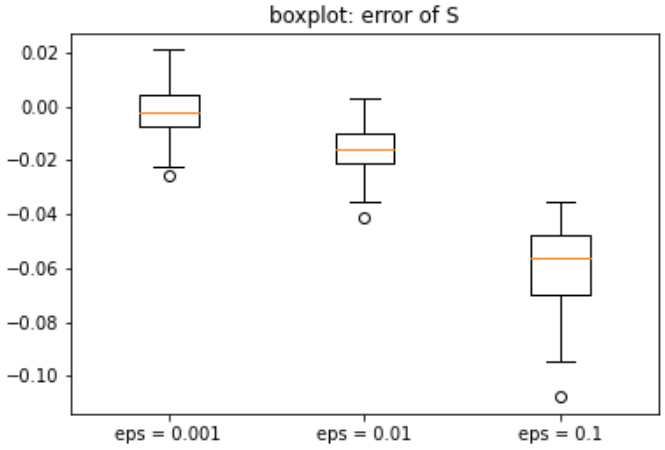
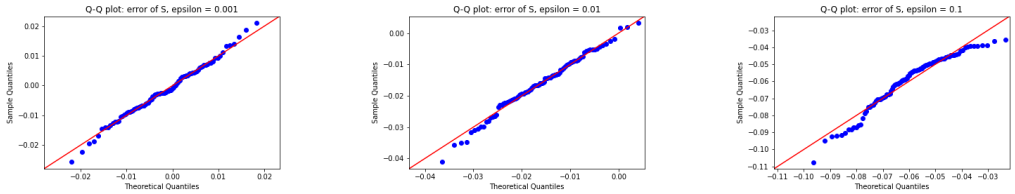
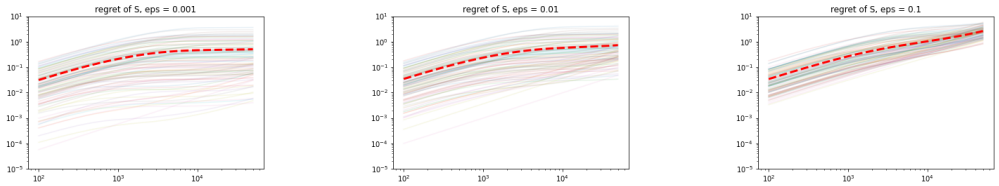


Figure 4.10: Boxplot of error of $\hat{S}_\varepsilon$



(a) $\varepsilon = 0.001$ (b) $\varepsilon = 0.01$ (c) $\varepsilon = 0.1$

Figure 4.11: qqplot of error of $\hat{S}_\varepsilon$

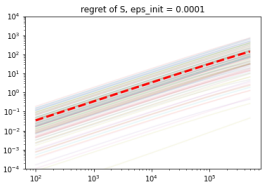


(a) $\varepsilon = 0.001$ (b) $\varepsilon = 0.01$ (c) $\varepsilon = 0.1$

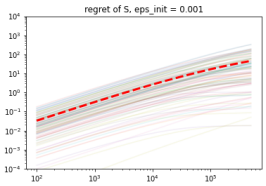
Figure 4.12: Regret of $\hat{S}_\varepsilon$

4.3.3 Calculation of $\hat{S}_{\varepsilon_t}(\mu, \nu)$ with decreasing ε_t

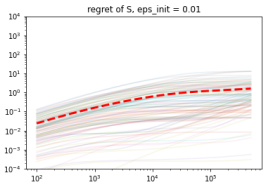
The calculation process and the range of initial ϵ_0 are same as the setting in 4.2.3. The results are shown in 4.13.



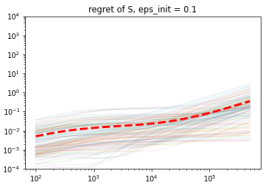
(a) $\varepsilon_0 = 0.0001$



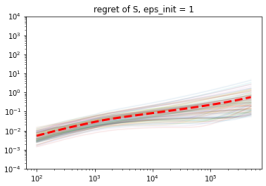
(b) $\varepsilon_0 = 0.001$



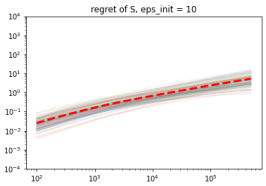
(c) $\varepsilon_0 = 0.01$



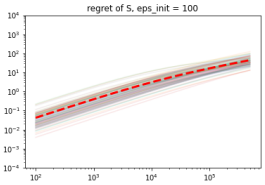
(d) $\varepsilon_0 = 0.1$



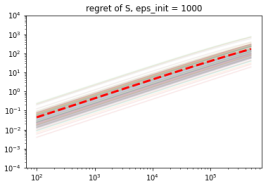
(e) $\varepsilon_0 = 1$



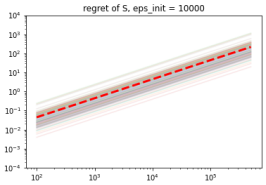
(f) $\varepsilon_0 = 10$



(g) $\varepsilon_0 = 100$



(h) $\varepsilon_0 = 1000$



(i) $\varepsilon_0 = 10000$

Figure 4.13: Regret of $\hat{S}_{\varepsilon_t}$, semi-discrete case

Chapter 5

Conclusion

In this paper, we utilize two stochastic algorithms, SAG (1) and ASGD (2), to approximate the squared Wasserstein distance W_2^2 in combination of two ideas: 1. adopting the empirical Sinkhorn divergence as the estimator of W_2^2, since it has the debiasing terms so that Sinkhorn divergence between two same measures is 0; 2. a regularization level ε that decreases with iteration, which takes advantage of the fact that at the beginning of the iterations, a larger ε gives better convexity to our objective function and makes convergence easier, while at the end of the iterations, a smaller ε gives better precision to our objective function.

Following these ideas, we first performed a theoretical analysis on the regret bound, which showed that the regret bound of empirical Sinkhorn divergence by using SAG could not avoid a large order with respect to T, where T is the number of iteration, especially when ε is decreasing along the iterations. In the other hand, the regret bound of regularized Wasserstein distance by using ASGD has a smaller order of T.

Then we performed the numerical experiments. First of all, we can see that the error in estimating W_2^2 directly using ASGD is not negligible. Afterwards, when setting the empirical sinkhorn divergence as the estimator, the SAG has a better performance, especially in the decreasing ε setting with $\varepsilon_0 = 1$. However, it is worth noting that results of the regret are heavily influenced by the choice of ε or ε_0.

The experimental results we obtained do not fit well with the theoretical analysis. In the theoretical analysis, ASGD has better performance, but SAG performed well in the experiments. This indicates that our theoretical analysis needs to be further optimized. First of all, the sensitivity to the initial ε in the experimental results can be the target of the next analysis. We can also increase the complexity of the assumptions on ε and n in the theoretical analysis.

Acknowledgments

I would like to thank my two supervisors, Olivier Wintenberger and Antoine Godichon-Baggioni for their guidance in thinking through many questions, showing me the direction when I was having difficulties, and for their advices on this report. I would also like to thank the other interns and PhDs in the LPSM lab for being so kind and making my five months so fulfilling and joyful.

Bibliography

[BBGS22] Bernard Bercu, Jérémie Bigot, Sébastien Gadat, and Emilia Siviero. A stochastic Gauss-Newton algorithm for regularized semi-discrete optimal transport. *arXiv:2107.05291 [math, stat]*, March 2022.

[CRL+20] Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster Wasserstein Distance Estimation with the Sinkhorn Divergence. *arXiv:2006.08172 [math, stat]*, October 2020.

[CTV19] Marco Cuturi, Olivier Teboul, and Jean-Philippe Vert. Differentiable Ranks and Sorting using Optimal Transport. page 11, 2019.

[Cut13] Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[DD20] Julie Delon and Agnès Desolneux. A Wasserstein-type distance in the space of Gaussian Mixture Models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.

[DGK18] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational Optimal Transport: Complexity by Accelerated Gradient Descent Is Better Than by Sinkhorn's Algorithm, June 2018.

[GCPB16] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic Optimization for Large-scale Optimal Transport. *arXiv:1605.08527 [cs, math]*, May 2016.

[GPC17] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning Generative Models with Sinkhorn Divergences. *arXiv:1706.00292 [stat]*, October 2017.

[Kan42] Leonid Kantorovich. On the transfer of masses. *Doklady Akademii Nauk USSR*, 1942.

[LCCC21] Xiucheng Li, Jin Yao Chin, Yile Chen, and Gao Cong. Sinkhorn Collaborative Filtering. In *Proceedings of the Web Conference 2021*, pages 582–592, Ljubljana Slovenia, April 2021. ACM.

[LYZZ19] Ruilin Li, Xiaojing Ye, Haomin Zhou, and Hongyuan Zha. Learning to Match via Inverse Optimal Transport. page 37, 2019.

[McD89] Colin McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference*, London Mathematical Society Lecture Note Series, pages 148–188. Cambridge University Press, Cambridge, 1989.

[MN19] Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[PJ92] B. T. Polyak and A. B. Juditsky. Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, July 1992.

[PW09] Ofir Pele and Michael Werman. Fast and robust Earth Mover's Distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467, September 2009.

[SRB13] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing Finite Sums with the Stochastic Average Gradient, 2013.

[SZ12] Ohad Shamir and Tong Zhang. Stochastic Gradient Descent for Non-smooth Optimization: Convergence Results and Optimal Averaging Schemes, December 2012.

[TBGS19] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein Auto-Encoders. *arXiv:1711.01558 [cs, stat]*, December 2019.