Stochastic Algorithm for Optimal Transport

Zehui Xuan

September 14th, 2022

Supervisor: Olivier Wintenberger, Antoine Godichon-Baggioni

Content

1. Introduction
2. Theoretical Analysis
   - Analysis on the upper bound of regret for the Sinkhorn Divergence
   - Analysis on the upper bound of regret for Regularized Wasserstein distance with fixed ε
3. Experiments
   - Calculation of the true value of the squared Wasserstein distance
   - Approximation in discrete case
   - Approximation in semi-discrete case
4. Conclusion

What is Optimal Transport ?

Optimal Transport:

- Moving one distribution of mass to another
- Minimizing the cost of transport

Example:

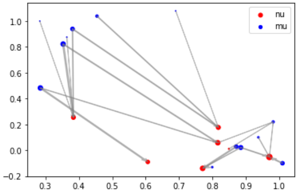


Figure 1: Example of OT in discrete case

Kantorovich formulation [4]:

$$W(\mu, \nu) = \min_{P \in U(\mu,\nu)} \sum_{i,j} c_{i,j} P_{i,j} \quad (1)$$

where
$U(\mu, \nu) = \{P \in \mathbb{R}_{+}^{I \times J} : P\mathbf{1}_J = \boldsymbol{\mu} \text{ and } P^\top \mathbf{1}_I = \boldsymbol{\nu}\}$
and c is a cost matrix.

Preliminaries (1/4)

Kantorovich formulation in continuous setting [4]:

$$\forall(\mu,\nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}), W(\mu,\nu) \overset{\text{def.}}{=} \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) \tag{2}$$

Preliminaries (1/4)

Kantorovich formulation in continuous setting [4]:

$$\forall(\mu,\nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}), W(\mu,\nu) \overset{\text{def.}}{=} \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) \tag{2}$$

Squared Wasserstein Distance [1]:

$$W_2^2(\mu,\nu) \overset{\text{def.}}{=} \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} \|x-y\|^2 \mathrm{d}\pi(x,y) \tag{3}$$

Preliminaries (1/4)

Kantorovich formulation in continuous setting [4]:

$$\forall(\mu,\nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{Y}), W(\mu,\nu) \overset{\text{def.}}{=} \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) \tag{2}$$

Squared Wasserstein Distance [1]:

$$W_2^2(\mu,\nu) \overset{\text{def.}}{=} \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} \|x-y\|^2\mathrm{d}\pi(x,y) \tag{3}$$

Regularized OT Cost [2]:

$$W_\varepsilon(\mu,\nu) \overset{\text{def.}}{=} \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon\mathrm{KL}(\pi \mid \mu\otimes\nu) \tag{4}$$

Preliminaries (2/4)

The semi-dual formulation in the discrete case [3]:

$$W_\varepsilon(\mu,\nu) = \max_{\mathbf{v}\in\mathbb{R}^J} \bar{H}_\varepsilon(\mathbf{v}) = \sum_{i=1}^{I} \bar{h}_\varepsilon(x_i,\mathbf{v})\boldsymbol{\mu}_i, \tag{5}$$

where

$$\bar{h}_\varepsilon(x,\mathbf{v}) = \sum_{j=1}^{J} \mathbf{v}_j\boldsymbol{\nu}_j + \begin{cases} -\varepsilon \ln\left(\sum_{j=1}^{J} \exp\left(\frac{\mathbf{v}_j - c\left(x,y_j\right)}{\varepsilon}\right)\boldsymbol{\nu}_j\right) - \varepsilon & \text{if } \varepsilon > 0, \\ \min_j\left(c\left(x,y_j\right) - \mathbf{v}_j\right) & \text{if } \varepsilon = 0. \end{cases} \tag{6}$$

Preliminaries (2/4)

The semi-dual formulation in the discrete case [3]:

$$W_\varepsilon(\mu,\nu) = \max_{\mathbf{v}\in\mathbb{R}^J} \bar{H}_\varepsilon(\mathbf{v}) = \sum_{i=1}^{I} \bar{h}_\varepsilon(x_i,\mathbf{v})\boldsymbol{\mu}_i, \tag{5}$$

where

$$\bar{h}_\varepsilon(x,\mathbf{v}) = \sum_{j=1}^{J} \mathbf{v}_j\boldsymbol{\nu}_j + \begin{cases} -\varepsilon \ln\left(\sum_{j=1}^{J} \exp\left(\frac{\mathbf{v}_j - c(x,y_j)}{\varepsilon}\right)\boldsymbol{\nu}_j\right) - \varepsilon & \text{if } \varepsilon > 0, \\ \min_j \left(c\left(x,y_j\right) - \mathbf{v}_j\right) & \text{if } \varepsilon = 0. \end{cases} \tag{6}$$

In the case of semi-discrete, the semi-dual problem is in a form of maximization of expectation as follows [3].

$$W_\varepsilon(\mu,\nu) = \max_{\mathbf{v}\in\mathbb{R}^J} \mathbb{E}_X\left[\bar{h}_\varepsilon(X,\mathbf{v})\right] \tag{7}$$

where $X \sim \mu$.

Preliminaries (3/4)

The gradient of function $\bar{h}_\varepsilon$:

$$\nabla_v \bar{h}_\varepsilon(x, \mathbf{v}) = \boldsymbol{\nu} - \chi^{\varepsilon}_{(c(x,y_\ell)-\mathbf{v}_\ell)_\ell}, \quad \text{where} \quad \forall \varepsilon > 0, (\chi^{\varepsilon}_{r})_j \stackrel{\text{def.}}{=} e^{-\frac{r_j}{\varepsilon}} \boldsymbol{\nu}_j \left(\sum_\ell e^{-\frac{r_\ell}{\varepsilon}} \boldsymbol{\nu}_\ell \right)^{-1}. \tag{8}$$

We mainly use two algorithms which utilize the gradient information.
Consider the discrete OT setting where $\mu = \sum_{i=1}^{I} \boldsymbol{\mu}_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{J} \boldsymbol{\nu}_j \delta_{y_j}$:

Algorithm 1 Stochastic Average Gradient (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** $k = 1, 2, \ldots$ **do**
  - Sample $i \in \{1, 2, \ldots, I\}$ 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**

Preliminaries (4/4)

Consider the discrete OT setting where μ is a continuous measure and $\nu = \sum_{j=1}^{J} \boldsymbol{\nu}_j \delta_{y_j}$:

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 $k = 1, 2, \ldots$ **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

Motivation

Why is it difficult to resolve an OT problem?

- Computational burden
- Large bias of the estimators

Motivation

Why is it difficult to resolve an OT problem?

- Computational burden
- Large bias of the estimators

Our goal: Estimating the squared Wasserstein distance in semi-discrete setting.

Motivation

Why is it difficult to resolve an OT problem?

- Computational burden
- Large bias of the estimators

Our goal: Estimating the squared Wasserstein distance in semi-discrete setting.

Our ideas:

1. We use the empirical Sinkhorn Divergence as our estimator.
   Sinkhorn Divergence:

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

2. We use a ε_t which decreases with iteration t, for example $\varepsilon_t = \frac{\varepsilon_0}{\sqrt{t}}$.

1. Introduction
2. Theoretical Analysis
   - Analysis on the upper bound of regret for the Sinkhorn Divergence
   - Analysis on the upper bound of regret for Regularized Wasserstein distance with fixed ε
3. Experiments
   - Calculation of the true value of the squared Wasserstein distance
   - Approximation in discrete case
   - Approximation in semi-discrete case
4. Conclusion

Contents

Analysis of the bounds for

- $\sum_{t=1}^{T} \left| \mathbb{E}\left[\hat{S}_{\varepsilon,n}^{(t)}\right] - W_2^2 \right|$, where $\hat{S}_{\varepsilon,n}^{(t)}$ is an approximation of the empirical Sinkhorn Divergence using SAG algorithm at iteration t;
- $\sum_{t=1}^{T} \left| \mathbb{E}\left[\hat{W}_{\varepsilon}^{(t)}\right] - W_2^2 \right|$, where $\hat{W}_{\varepsilon}^{(t)}$ is an approximation of regularized Wasserstein distance using ASGD algorithm at iteration t.

Decomposition

Consider the decomposition:

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

with the bound for each component:

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

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

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

with probability $1-\delta$, and finally

$$\left|\mathbb{E}\left[\hat{S}_{\varepsilon,n}^{(t)}\right] - \hat{S}_{\varepsilon,n}\right| \leq \frac{1}{t}\left(C_{11}n^{-1}\varepsilon^{-1} + C_{12}\varepsilon^{-1} + C_{21}n^2\varepsilon + C_{22}\varepsilon + C_{31}n + C_{32}\right)$$

where C_{11}, C_{12}, C_{21}, C_{22}, C_{31}, C_{32} and C_4 are constants. (Details in Appendix 2)

Fixed ε - Result

We have

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

where C_{11}, C_{12}, C_{21}, C_{22}, C_{31}, C_{32} and C_4 are constants. (Details in Appendix 2)

Fixed ε - Analysis (1/3)

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_T as follows.

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

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

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

Fixed ε - Analysis (2/3)

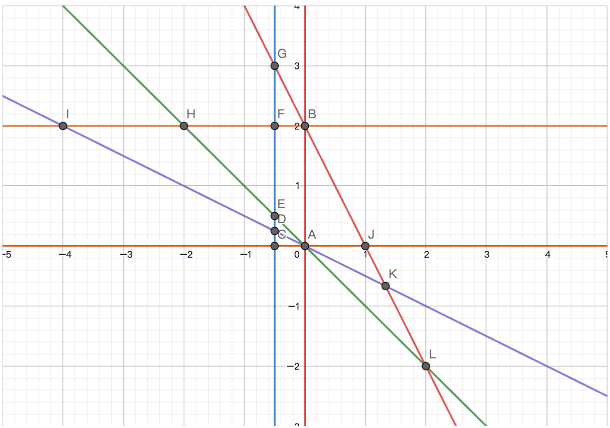


Figure 2: Illustration of the objective on $\mathbb{R}^2$

Fixed ε - Analysis (3/3)

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

Table 1: Analysis of order of T

Decreasing ε - Result

We have

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

where C_{11}, C_{12}, C_{21}, C_{22}, C_{31}, C_{32} and C_4 are constants. (Details in Appendix 2)

Decreasing ε - Analysis (1/2)

Let $n = AT^a$. We rewrite the bound B_T as follows.

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

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

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

Decreasing ε - Analysis (2/2)

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 2: Analysis of order of T

Decomposition

Consider the decomposition:

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

where

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

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

Result and analysis

We have

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

where

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

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 $O(T^{\frac{1}{2}} \ln T)$.

1. Introduction

2. Theoretical Analysis
    - Analysis on the upper bound of regret for the Sinkhorn Divergence
    - Analysis on the upper bound of regret for Regularized Wasserstein distance with fixed ε

3. Experiments
    - Calculation of the true value of the squared Wasserstein distance
    - Approximation in discrete case
    - Approximation in semi-discrete case

4. Conclusion

Contents

- Calculation of the true values of $W_2^2(\mu, \nu)$ which are used for comparison with experimental results.
- Approximation of $W_2^2(\mu, \nu)$ in discrete setting,
  - with fixed ε;
  - with decreasing ε_t.
- Approximation of $W_2^2(\mu, \nu)$ in semi-discrete setting,
  - with fixed ε;
  - with decreasing ε_t.

Methods

In decrease setting:

1. Solve the problem directly by using linear programming solver.
2. Calculate the total transport cost.

In semi-discrete setting:

1. Solve the regularized problem by using ASGD algorithm with $\varepsilon = 0$.
2. Discretize μ with sample size N.
3. Calculate the total transport cost.

Verification of the method in semi-discrete setting

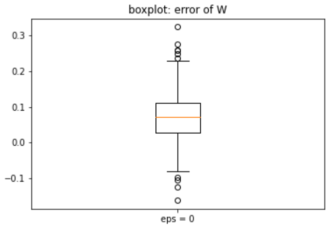


Figure 3: Boxplot of error of W

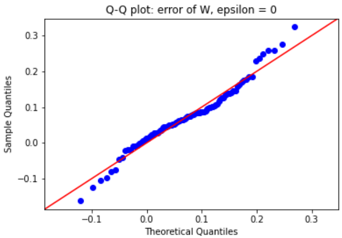
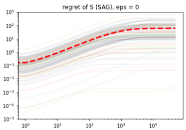
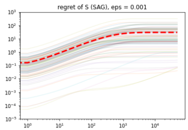


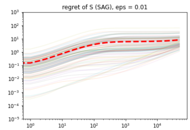
Figure 4: Q-Q plot of error of W



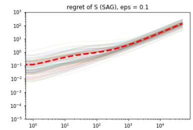
(a) $\varepsilon = 0$



(b) $\varepsilon = 0.001$

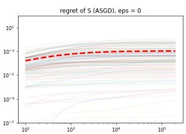


(c) $\varepsilon = 0.01$

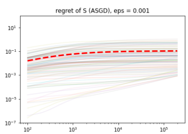


(d) $\varepsilon = 0.1$

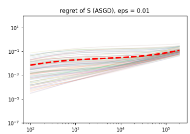
Figure 5: Regret of $\hat{S}_\varepsilon$ (SAG), discrete case



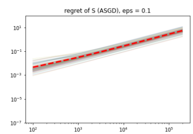
(a) $\varepsilon = 0$



(b) $\varepsilon = 0.001$



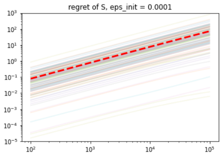
(c) $\varepsilon = 0.01$



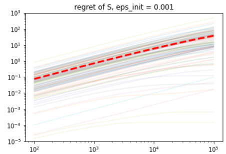
(d) $\varepsilon = 0.1$

Figure 6: "Regret" of $\hat{S}_\varepsilon$ (ASGD), discrete case

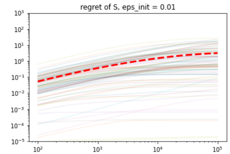
Approximation in discrete case



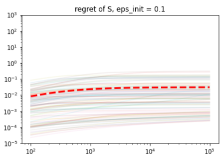
(a) $\varepsilon_0 = 0.0001$



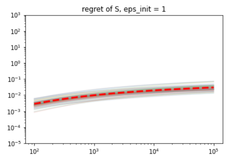
(b) $\varepsilon_0 = 0.001$



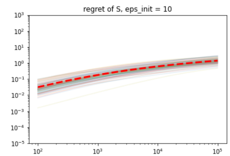
(c) $\varepsilon_0 = 0.01$



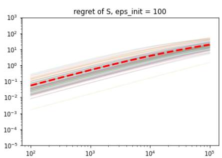
(d) $\varepsilon_0 = 0.1$



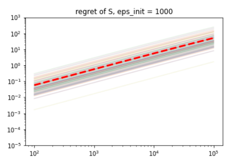
(e) $\varepsilon_0 = 1$



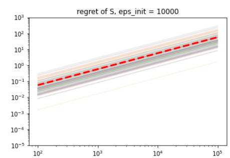
(f) $\varepsilon_0 = 10$



(g) $\varepsilon_0 = 100$



(h) $\varepsilon_0 = 1000$



(i) $\varepsilon_0 = 10000$

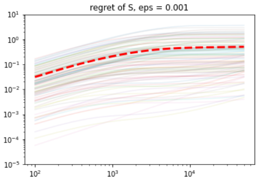
Figure 7: "Regret" of $\hat{S}_{\varepsilon_t}$, discrete case

1. Introduction

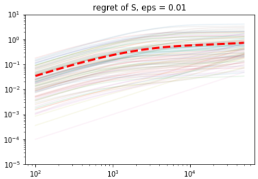
2. Theoretical Analysis
   - Analysis on the upper bound of regret for the Sinkhorn Divergence
   - Analysis on the upper bound of regret for Regularized Wasserstein distance with fixed ε

3. Experiments
   - Calculation of the true value of the squared Wasserstein distance
   - Approximation in discrete case
   - Approximation in semi-discrete case

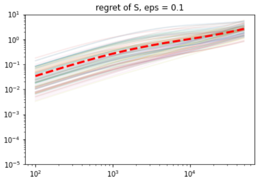
4. Conclusion



(a) $\varepsilon = 0.001$



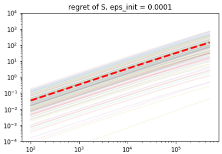
(b) $\varepsilon = 0.01$



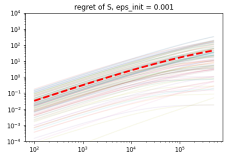
(c) $\varepsilon = 0.1$

Figure 8: "Regret" of $\hat{S}_\varepsilon$, semi-discrete case

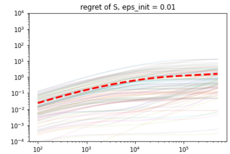
Approximation in semi-discrete case



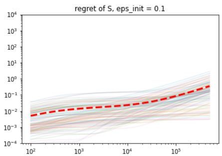
(a) $\varepsilon_0 = 0.0001$



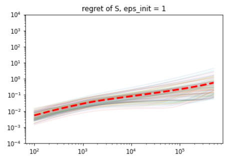
(b) $\varepsilon_0 = 0.001$



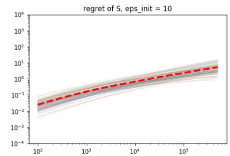
(c) $\varepsilon_0 = 0.01$



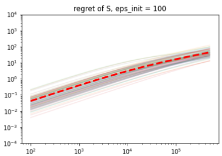
(d) $\varepsilon_0 = 0.1$



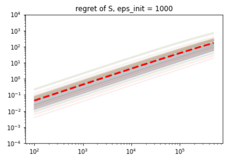
(e) $\varepsilon_0 = 1$



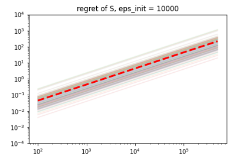
(f) $\varepsilon_0 = 10$



(g) $\varepsilon_0 = 100$



(h) $\varepsilon_0 = 1000$



(i) $\varepsilon_0 = 10000$

Figure 9: "Regret" of $\hat{S}_{\varepsilon_t}$, semi-discrete case

1. Introduction
2. Theoretical Analysis
   - Analysis on the upper bound of regret for the Sinkhorn Divergence
   - Analysis on the upper bound of regret for Regularized Wasserstein distance with fixed ε
3. Experiments
   - Calculation of the true value of the squared Wasserstein distance
   - Approximation in discrete case
   - Approximation in semi-discrete case
4. Conclusion

Conclusion

- Theoretical analysis on the upper bound of regret:
  - The upper bound of regret of the empirical Sinkhorn divergence by using SAG could not avoid a large order with respect to T.
  - The upper bound of regret of the regularized Wasserstein distance by using ASGD has a smaller order of T.
- Numerical experiments:
  - Error in estimating W_2^2 directly using ASGD is not negligible.
  - Results for SAG is heavily influenced by the choice of ε or ε_0.

Thank you

Thank you for listening !

Bibliography

[1] 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.

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

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

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

5. Appendix
    - Semi-dual formulation of regularized OT in continuous setting
    - List of constants in the regret analysis for the Sinkhorn Divergence

5. Appendix
   - Semi-dual formulation of regularized OT in continuous setting
   - List of constants in the regret analysis for the Sinkhorn Divergence

Semi-dual formulation of regularized OT in continuous setting

The semi-dual formulation of regularized OT problem [3]:

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

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

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

5 Appendix
- Semi-dual formulation of regularized OT in continuous setting
- List of constants in the regret analysis for the Sinkhorn Divergence

List of constants in the regret analysis for the Sinkhorn Divergence

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

List of constants in the regret analysis for the Sinkhorn Divergence

$$W_\varepsilon(\hat{\mu}_n, \nu_J) = \max_{\mathbf{v}_1 \in \mathbb{R}^J} \bar{H}_\varepsilon^1(\mathbf{v}_1) = \frac{1}{n}\sum_{i=1}^{n} \bar{h}_\varepsilon^1(x_i, \mathbf{v}_1),$$

$$\bar{h}_\varepsilon^1(x, \mathbf{v}) = \sum_{j=1}^{J} \mathbf{v}_j \boldsymbol{\nu}_j - \varepsilon \ln\left(\sum_{j=1}^{J} \exp\left(\frac{\mathbf{v}_j - c(x, y_j)}{\varepsilon}\right)\boldsymbol{\nu}_j\right) - \varepsilon,$$

$$W_\varepsilon(\hat{\mu}_n, \hat{\mu}_n) = \max_{\mathbf{v}_2 \in \mathbb{R}^n} \bar{H}_\varepsilon^2(\mathbf{v}_2) = \frac{1}{n}\sum_{i=1}^{n} \bar{h}_\varepsilon^2(x_i, \mathbf{v}_2),$$

$$\bar{h}_\varepsilon^2(x, \mathbf{v}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{v}_i - \varepsilon \ln\left(\frac{1}{n}\sum_{i=1}^{n} \exp\left(\frac{\mathbf{v}_i - c(x, x_i)}{\varepsilon}\right)\right) - \varepsilon,$$

$$W_\varepsilon(\nu_J, \nu_J) = \max_{\mathbf{v}_3 \in \mathbb{R}^J} \bar{H}_\varepsilon^3(\mathbf{v}_3) = \sum_{j=1}^{J} \bar{h}_\varepsilon^3(y_j, \mathbf{v}_3)\boldsymbol{\nu}_j,$$

$$\bar{h}_\varepsilon^3(y, \mathbf{v}) = \sum_{j=1}^{J} \mathbf{v}_j \boldsymbol{\nu}_j - \varepsilon \ln\left(\sum_{j=1}^{J} \exp\left(\frac{\mathbf{v}_j - c(y, y_j)}{\varepsilon}\right)\boldsymbol{\nu}_j\right) - \varepsilon.$$

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

where

$$
\left|\mathbf{E}\left[\bar{H}_\varepsilon^1(\mathbf{v}_1^{(t)})\right] - \bar{H}_\varepsilon^1(\mathbf{v}_1^*)\right| \leq \frac{32n}{t}\left(\bar{H}_\varepsilon^1(\mathbf{v}_1^*) - \bar{H}_\varepsilon^1(\mathbf{v}_1^0) + \frac{4}{n^2\varepsilon}\|\mathbf{v}_1^0 - \mathbf{v}_1^*\|^2 + \frac{\sigma_1^2 n\varepsilon}{16}\right)
$$

$$
\frac{1}{2}\left|\mathbf{E}\left[\bar{H}_\varepsilon^2(\mathbf{v}_2^{(t)})\right] - \bar{H}_\varepsilon^2(\mathbf{v}_2^*)\right| \leq \frac{16n}{t}\left(\bar{H}_\varepsilon^2(\mathbf{v}_2^*) - \bar{H}_\varepsilon^2(\mathbf{v}_2^0) + \frac{4}{n^2\varepsilon}\|\mathbf{v}_2^0 - \mathbf{v}_2^*\|^2 + \frac{\sigma_2^2 n\varepsilon}{16}\right)
$$

$$
\frac{1}{2}\left|\mathbf{E}\left[\bar{H}_\varepsilon^3(\mathbf{v}_3^{(t)})\right] - \bar{H}_\varepsilon^3(\mathbf{v}_3^*)\right| \leq \frac{16J}{t}\left(\bar{H}_\varepsilon^3(\mathbf{v}_3^*) - \bar{H}_\varepsilon^3(\mathbf{v}_3^0) + \frac{4}{J\varepsilon}\|\mathbf{v}_3^0 - \mathbf{v}_3^*\|^2 + \frac{\sigma_3^2 \varepsilon}{16}\right)
$$