Optimal Transport Project Report

Stochastic Optimization for Large-scale Optimal Transport

Eugène Berta
MVA, ENS Paris Saclay
Télécom Paris
eugene.berta@gmail.com

January 2023

Abstract

The paper we study tackles the problem of computing the Optimal Transport (OT) cost in different settings. It introduces new algorithms, relying on the Dual and Semi-dual reformulation of the original Kantorovich problem to compute OT between *(i)* two discrete distributions (discrete case) *(ii)* a discrete distribution and a continuous distribution (semi-discrete case) *(iii)* two continuous distributions (continuous case). The originality of the paper is to formulate OT as an expectation maximization problem, leveraging existing stochastic algorithms to compute the cost in each of the three settings. In the discrete and semi-discrete cases, the algorithms proposed provide faster convergence speed than previous tools used in the literature. The article introduces the first proposed method to compute OT in the continuous setting. Theses new methods rely on the ability to sample from the continuous distributions involved. This is relevant because, *(a)* sampling from a continuous distribution is often easier than discretizing it on a grid to get back to the discrete case *(b)* the algorithms free themselves from the discretization error induced by this naive technique. For each of these settings, we will first expose the method proposed in the paper, trying to make as clear as possible the derivation from the initial OT problem to the stochastic algorithm associated. Then, we will propose an implementation of the algorithm, run experiments and present the results obtained. Finally, we will discuss the limitations of the paper, make a few personal remarks and evoke possible extensions.

1 Introduction

The paper studied [1] is part of a wider research effort to develop geometrically faithful ways to compare probability distributions. In machine learning applications, Kullback-Leibler (KL) divergence is the standard way to compute a "distance" between probability distributions $\alpha \in \mathcal{M}_1^+(\mathcal{X})$ and $\beta \in \mathcal{M}_1^+(\mathcal{Y})$. KL has a natural connection to Maximum Likelihood Estimator in Statistics and comes with many interesting properties. However, it is blind to the geometry of the support $\mathcal{X}$. It is thus said to be a non-geometric divergence and in particular, it does not metrize the convergence in law. For some applications, the geometry of the support is important to the problem, the most emblematic example in Machine Learning being the training of Generative Adversarial Network [2], [3]. In these applications, using KL is not a satisfying solution.

Optimal Transport is a class of geometric divergences that is gaining traction in data science. OT directly lifts a distance on the feature space $\mathcal{X} \times \mathcal{Y}$ to build a distance on the probability measure space $\mathcal{M}_1^+(\mathcal{X}) \times \mathcal{M}_1^+(\mathcal{Y})$:

$$OT(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\pi_1=\alpha, \pi_2=\beta} \int_{\mathcal{X}\times\mathcal{Y}} C d\pi \tag{1}$$

where $C(x, y)$ is a symmetric positive cost function (in practice we often use $||x - y||^2$ which induces the 2-Wasserstein distance on the probability measures space).

OT is geometrical by construction and it metrizes the convergence in law. Moreover, OT comes with very strong theoretical guarantees. However, solving the OT problem in the general case is very computationally intensive. In practice, this is the main limitation to a wider use of OT.

Regularized Optimal Transport In 2013, [4] introduced a regularized version of the original cost that made possible the use of the GPU compatible *Sinkhorn algorithm* [5] to solve the problem at scale:

$$OT_\epsilon(\alpha, \beta) \stackrel{\text{def.}}{=} \min_{\pi_1=\alpha, \pi_2=\beta} \int_{\mathcal{X}\times\mathcal{Y}} C d\pi + \epsilon KL(\pi, \alpha \otimes \beta) \tag{2}$$

Regularized OT lifted the computational burden of OT and thus helped democratize its applications. However, it is clear from the equation above that the optimal solution is different from the solution of the original problem. The regularization term introduces an **entropic bias** in the solution computed.

Unlike the classical OT solution, the regularized OT minimizer does not cover the full support of the target distribution. This shows that $OT\epsilon$ loses in part the geometric advantages of OT. Sinkhorn divergences [6] have later been introduced to mitigate the entropic bias in regularized optimal transport while keeping the computational advantages of the regularized problem (2).

The dual OT problems. Using the formulation $\langle f, \alpha \rangle = \int f d\alpha$, the classical OT problem (1) can be seen as a linear optimisation problem. It's dual reads:

$$OT(\alpha, \beta) = \max_{f,g \in \mathcal{R}(\mathcal{C})} \int_{\mathcal{X}} f d\alpha + \int_{\mathcal{Y}} g d\beta \tag{3}$$

With,

$$\mathcal{R}(\mathcal{C}) = \{(f, g) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}), f(x) + g(y) - C(x, y) \leq 0, \forall (x, y) \in \mathcal{X} \times \mathcal{Y}\}$$

The regularized OT problem (2) is ϵ strongly convex, for $\epsilon > 0$ its dual reads:

$$OT_\epsilon(\alpha, \beta) = \max_{f \in \mathcal{C}(\mathcal{X}), g \in \mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} f d\alpha + \int_{\mathcal{Y}} g d\beta - \epsilon \int_{\mathcal{X} \times \mathcal{Y}} exp\Big(\frac{f + g - C}{\epsilon}\Big) d\alpha d\beta \tag{4}$$

We refer to [7] for more details.

C-transforms and semi dual formulations. We define the c-transform f^c as:

$$\forall y \in \mathcal{Y}, f^c(y) \stackrel{\text{def.}}{=} \min_{x \in \mathcal{X}} C(x, y) - f(x)$$

Using this expression, we can re-formulate (3) as :

$$OT(\alpha, \beta) = \max_{f \in \mathcal{C}(\mathcal{X})} \int_{\mathcal{X}} f d\alpha + \int_{\mathcal{Y}} f^c d\beta \tag{5}$$

We call this problem the semi-dual of (1).

Similarly, we define the ϵ-c-transform $f^{c,\epsilon}$ as in [7] (5.3 Entropic Semidiscrete Formulation):

$$\forall y \in \mathcal{Y}, f^{c,\epsilon}(y) \stackrel{\text{def.}}{=} -\epsilon \log \left(\int_{\mathcal{X}} exp\Big(\frac{-C(x, y) + f(x)}{\epsilon}\Big) d\alpha(x) \right)$$

This allows us to derive the semi-dual of (2):

$$OT_\epsilon(\alpha, \beta) = \max_{f \in \mathcal{C}(\mathcal{X})} \int_{\mathcal{X}} f d\alpha + \int_{\mathcal{Y}} f^{c,\epsilon} d\beta - \epsilon \tag{6}$$

Optimal Transport as an Expectation maximisation problem. The first important contribution of the paper, that is the foundation to all the algorithms introduced, is to reformulate these dual and semi-dual as expectation maximisation problems :

$$\begin{aligned} OT(\alpha, \beta) &= \max_{f,g \in \mathcal{R}(\mathcal{C})} \mathbb{E}_{X,Y}\big[f(x) + g(y)\big] \\ &= \max_{f \in \mathcal{C}(\mathcal{X})} \mathbb{E}_Y\big[\int_{\mathcal{X}} f d\alpha + f^c(y)\big] \end{aligned} \tag{7}$$

$$\begin{aligned} OT_\epsilon(\alpha, \beta) &= \max_{f,g \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y})} \mathbb{E}_{X,Y}\big[f(x) + g(y) - \epsilon exp\Big(\frac{f(x) + g(y) - C(x,y)}{\epsilon}\Big)\big] \\ &= \max_{f \in \mathcal{C}(\mathcal{X})} \mathbb{E}_Y\big[\int_{\mathcal{X}} f d\alpha + f^{c,\epsilon}(y) - \epsilon\big] \end{aligned} \tag{8}$$

This trick is at the heart of the studied paper. Indeed, viewing the problem through the lens of expectation maximisation enables to leverage stochastic algorithms to solve the Optimal Transport problem, as we will expose in more details in the following parts.

2 Discrete Optimal Transport

We will first consider the case were α and β are discrete distributions, that is $\alpha = \sum_{i=1}^n \alpha_i \delta_{x_i}$ and $\beta = \sum_{j=1}^m \beta_j \delta_{y_j}$.

2.1 Method and algorithm

In the discrete case, we can re-write the expectation maximization problem (8):

$$\begin{aligned} OT_\epsilon(\alpha, \beta) &= \max_{f \in \mathbb{R}^n} \mathbb{E}_Y\big[\sum_{i=1}^n f_i \alpha_i - \epsilon \log\left(\sum_{i=1}^n exp\Big(\frac{f_i - C(x_i, y_j)}{\epsilon}\Big)\alpha_i\right) - \epsilon\big] \\ &= \max_{f \in \mathbb{R}^n} \sum_{j=1}^m \left[\sum_{i=1}^n f_i \alpha_i - \epsilon \log\left(\sum_{i=1}^n exp\Big(\frac{f_i - C(x_i, y_j)}{\epsilon}\Big)\alpha_i\right) - \epsilon\right]\beta_i \end{aligned}$$

We can use Stochastic Gradient Descent (SGD) to maximize this sum. At each iteration, we pick an index $k \in [\![1, m]\!]$ with probability β_k, and we

perform an update $f = f + \lambda \nabla_{f_k} H(f, x)$ where λ is a carefully chosen step size and:

$$H(f,x) = \sum_{i=1}^{n} f_i \alpha_i - \epsilon \log \left(\sum_{i=1}^{n} exp\Big(\frac{f_i - C(x_i, y_j)}{\epsilon}\Big) \alpha_i \right) - \epsilon$$

Thus,

$$\nabla_{f_k} H(f,x) = \alpha_k - \epsilon exp\left(\frac{f_k - C(x_k, y_j)}{\epsilon}\right) \alpha_k \left(\sum_{i=1}^{n} exp\Big(\frac{f_i - C(x_i, y_j)}{\epsilon}\Big) \alpha_i \right)^{-1}$$

This is a closed form expression for our gradient, we can apply Stochastic Gradient Descent.

Noticing that we sample in a finite distribution, the author propose to fasten the computation using Stochastic Averaged Gradient (SAG) [8] instead of SGD. We refer to the original paper for details but this is a simple way to get a linear convergence rate $\mathcal{O}(1/k)$ instead of the $\mathcal{O}(1/\sqrt{k})$ using SGD.

After convergence, this algorithm gives us an optimal f solving (6), we recover the associated optimal g by $g = f^{c,\epsilon}$. As explained in [9], these regularized dual potentials are linked to (u, v) solutions of Sinkhorn algorithm by $(u, v) = (\alpha_i exp(f/\epsilon), \beta_j exp(g/\epsilon))$. So, we can recover the optimal coupling:

$$P_{i,j} = exp\Big(\frac{f(x_i) + g(y_j) - C(x_i, y_j)}{\epsilon}\Big) \alpha_i \beta_j$$

For the unregularized case ($\epsilon = 0$) the expression of the gradient we derived is not valid. We refer to the original paper for the gradient update in this case.

2.2 Numerics

Classical Optimal Transport. To test the algorithm with $\epsilon = 0$ we implement :

- A true OT solver using the convex optimization package cvxpy [10].
- The SAG introduced in the paper with the adequate modification for $\epsilon = 0$, we call it "SAG-OT". Our implementation is made available in the notebook that accompanies this report (see 6).

We compare the results obtained by these two methods to validate our implementation on two distributions pairs visible on Figure 1 and Figure 2.

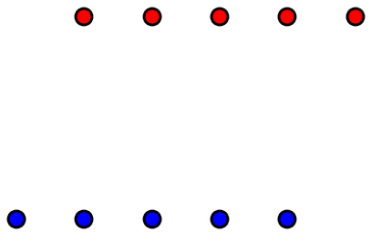

Figure 1: A very simple discrete distribution for which we know the optimal transport.

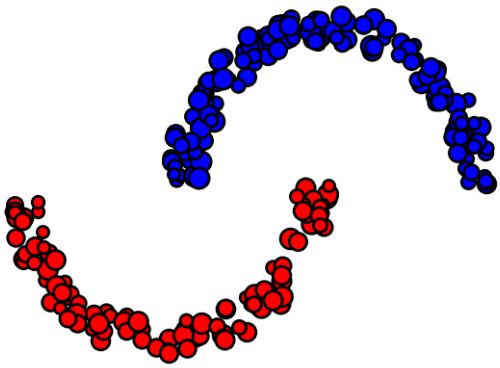

Figure 2: A more complex discrete distribution.

We remark that while the convex optimization solver solves the primal (1) and thus returns an optimal transport plan P^*, the algorithm proposed in the paper solves the dual (3) and returns optimal dual potentials (f^*, g^*). When $\epsilon = 0$, it is not obvious how to recover P^* from (f^*, g^*), thus, we limit our comparison to the distance computed by the two methods.

For the first distribution, both methods find the same cost $OT(\alpha, \beta) = 0.5$ with good numerical precision. For the second example however, we observe a small difference. $OT(\alpha, \beta) = 2.3257$ and $SAG - OT(\alpha, \beta) = 2.3199$ (with step size = .01 for 50000 iterations). Moreover, we observe that the convergence of SAG-OT is very dependant on the step size chosen.

Regularized Optimal Transport. To test the algorithm with $\epsilon > 0$ we implement :

- The Sinkhorn algorithm for regularized Optimal Transport.

- The SAG introduced in the paper for $\epsilon > 0$, we call it "SAG-OT_ϵ".

We use the same two distributions as before for the comparison, we set $\epsilon = 0.01$, we run Sinkhorn for 10000 iterations and SAG-OT_ϵ for 20000 iterations. For the toy distribution, we refer to the notebook, but the two algorithms find the optimal solution easily.

For the more complex example, the two algorithms converge (Figure 3 and Figure 4) to the same result. Visually, we see that the algorithms converge and we obtain the same transport plan with a good numerical precision.

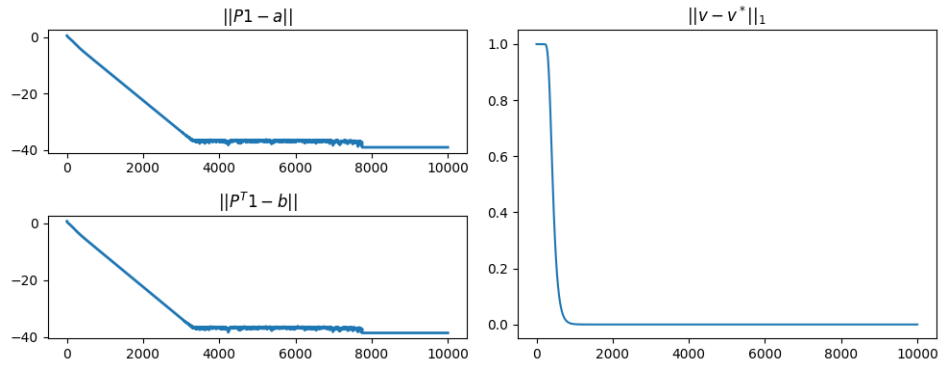

Figure 3: Convergence of the Sinkhorn algorithm on the complex example.

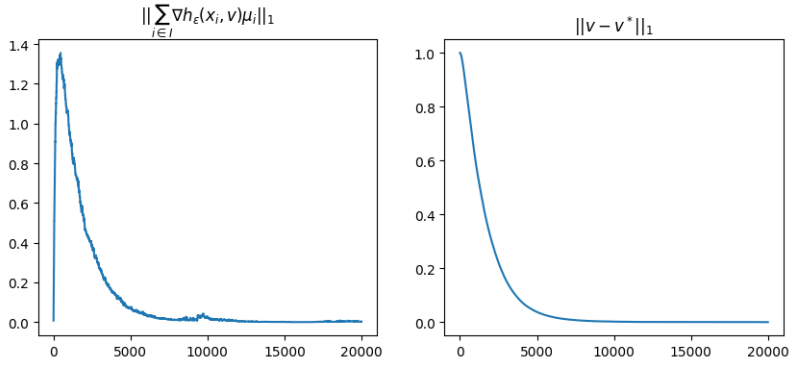

Figure 4: Convergence of SAG-OT_ϵ on the complex example.

In Figure 5 we plot the transport plans obtained for the two methods, along with the true transport plan P^* obtained by solving the primal problem with $\epsilon = 0$. Empirically, we observe that the regularized solution is less sparse

than the true solution (fewer sharp lines), this is the effect of the entropic regularization introduced.

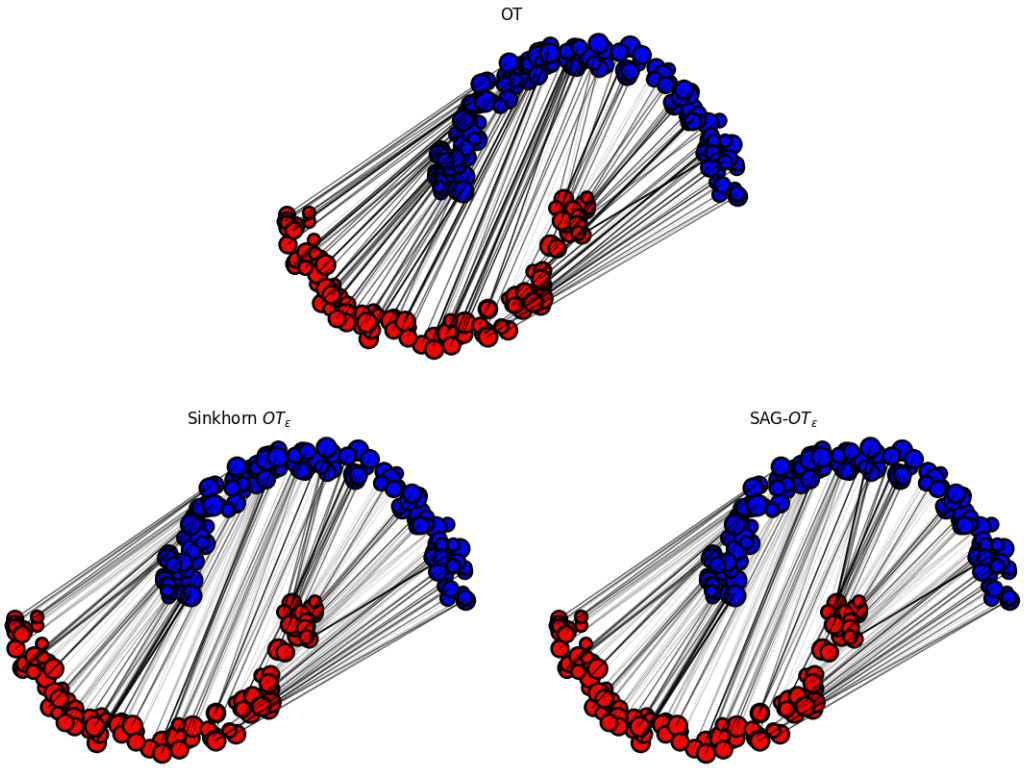

Figure 5: Transport plans for the three different methods.

2.3 Limitations

Our opinion is that the stochastic algorithm introduced for the discrete case is the most important contribution of the paper. We will further justify this claim later in the report, but improving the speed of the Sinkhorn algorithm in the discrete case has a direct impact on most of the known OT use cases. However, this improvement comes at the cost of setting the value for an additional parameter: the step size. This might mitigate the impact of this algorithm. Indeed, we can assume that most users of OT (at least for data science applications) are not specialist and will rely on an external library for the computation of the cost. For such users it might be worth it to sacrifice some computation speed to avoid having to find the right step size.

Finally, SAG is faster than SGD at the cost of a larger memory footprint, in real-world application, with millions of samples in high dimension, this might become problematic. We can mitigate the memory footprint of SAG using mini-batches instead of individual samples.

3 Semi-discrete Optimal Transport

In this part, we work with a discrete distribution $\alpha = \sum_{i=1}^{n} \alpha_i \delta_{x_i}$, and a continuous distribution β from which we can sample.

3.1 Method and algorithm

The key idea used in the paper to solve semi-discrete Optimal Transport is to work with the semi-dual (6). Indeed, if we choose β to be continuous and α discrete (the distance being symmetric this choice has no impact), the expectation maximization problem (8) is finite dimensional. This allows us to use a very similar approach to the previous section, the only difference being that we now sample in a continuous distribution in the SGD. The gradient update that we computed previously is still valid. However, we cannot use SAG anymore. There is a infinite number of different samples that we can draw from β so we cannot save the gradients in a table and update them regularly. To improve the convergence speed of the method, the authors propose the use of an Averaged SGD instead of a classical SGD, this allows to reach a convergence rate of $\mathcal{O}(1/\sqrt{k})$.

3.2 Numerics

We implement the Averaged SGD proposed for the case $\epsilon > 0$. To validate our implementation, we run it on the semi-discrete setting of Figure 6.

We use $\epsilon = .01$ and we run the Average SGD for 10000 iterations with a step size set to 0.1. The algorithm returns an optimal semi-dual potential f^* of size n.

To visualize our result, we draw 200 samples from the continuous distribution β, we assign a uniform weight to each sample.

- First, we compute the semi-discrete optimal transport P^*_{sd} using $f^{c,\epsilon}$ to compute g^*.
- Then, we compute the discrete optimal transport P^*_d using SAG-OT_ϵ.

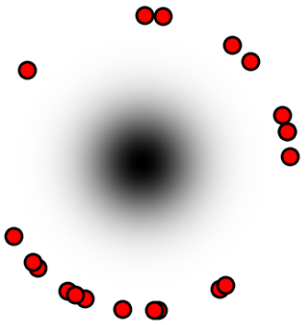

Figure 6: Semi discrete distribution used for our experiment.

Even-though the two transport plans are similar, we still observe some differences in Figure 7. This is not surprising. Indeed the semi-discrete transport plan corresponds to an idealized version of the discrete transport plan, where α is continuous. On the other hand, we use only 200 samples to build our discrete distribution. The approximation of the continuous distribution is pretty bad with this few samples. This illustrates the discretization error that semi-discrete and continuous OT try to get rid of.

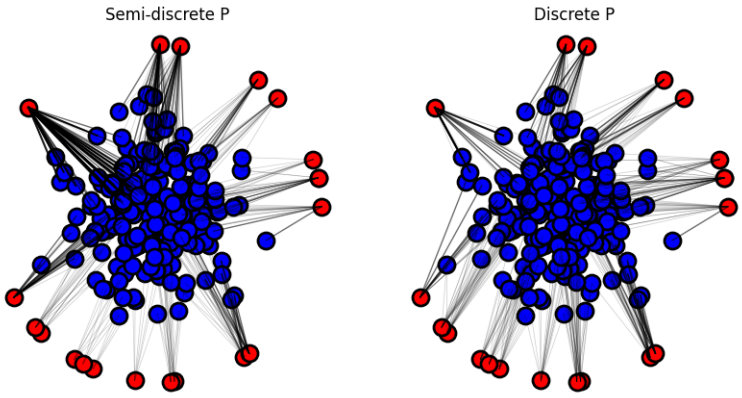

Figure 7: Comparison between the semi-discrete transport plan and the discrete transport plan after discretization.

3.3 Limitations

Our opinion is that there are very few applications where the use of semi-discrete optimal transport is justified. As we saw previously, if we can sample in the continuous distribution β, a convenient way to discretize it is to draw many samples with the same associated probability β_j rather than discretizing it on a grid. At the time of the paper, applications of Optimal Transport to data science were stammering, but modern discrete solvers can scale to millions of samples easily, drastically mitigating the issue of the discretization error. Moreover, in data science, datasets are very rarely continuous probability distributions but can generally be seen as the discretization of a latent continuous distribution. This argues that discrete optimal transport is of more crucial importance than semi-discrete and continuous settings.

4 Continuous Optimal Transport

In this part, we work with α and β, two continuous distributions from which we can sample.

4.1 Method and algorithm

In this case, the Expectation Maximization formulation of OT (8, line 2) cannot be written as the maximization over a finite vector f, the problem is infinite dimensional. We work with the initial dual problem (4) and it's Expectation Maximization version (8, line 1). When, $\epsilon = 0$, the constraints on the dual variable $f \oplus g \leq C$ prevents us from writing the problem as an Expectation Maximisation. Thus in this case, the provided algorithm only works for $\epsilon > 0$.

The main trick used here is to optimize over sums of kernels that approximate f and g: $f = \sum_{i=1}^{k} \alpha_i k(x_i, .)$ and $g = \sum_{i=1}^{k} \alpha_i k(y_i, .)$. This allows to re-write (8, line 1) as a finite dimension expectation maximization problem, on which we can use a SGD. The paper derive the update of that SGD and deduce a closed form formula for α_i. At each step of the algorithm, we compute a new term α_k and the approximation is getting better.

4.2 Numerics

We implement the proposed Kernel-SGD. To validate our implementation, we run our algorithm on a simple setting. It is hard to visualize the optimal

transport between continuous distributions, thus, we rely on the distance computed to illustrate our algorithm. We set ourselves in a very simple 1D setting, for which we can compute the Wasserstein distance in closed form. We define :

- A target normal distribution β with mean $\mu_\beta = 2$ and variance $\sigma_\beta = 0.2$.
- A model distribution α_θ with mean $\mu_\alpha = 4\theta$ and variance $\sigma_\alpha = 0.2$.

During the experiment, we linearly interpolate θ from 0 to 1 and compute the true Wasserstein distance $W_2(\alpha_\theta, \beta)$ and the distance computed by our Kernel-SGD $KSGD(\alpha_\theta, \beta)$. The setting of the experiment is illustrated in Figure 8.

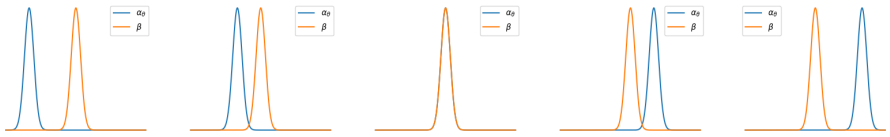

Figure 8: Continuous experiment setting.

We use the kernel $k(x, x') = exp(\frac{-||x-x'||}{\sigma^2})$ with $\sigma = 0.2$, we set $\epsilon = 0.1$ and we run the Kernel SGD for 10000 iterations. The result of the experiment is exposed in Figure 9.

We see that the distance is well approximated when the Gaussians are close to each other. When the distance augments however, the approximation is getting worse. Moreover, at some point, we see a the Kernel-SGD distance reaches a saturation point where the distance stays constant while the distributions are still getting further away from each other. This is due to the variance parameter chosen in the kernel. This illustrates well that the algorithm is very dependant on a correct tuning of the Kernel chosen.

4.3 Limitations

As exposed by the result of our experiment, the method proposed comes with inherent limitations. First, the computation is very lengthy even for the simple case of 1D gaussians, and a good approximation requires at least 10000 points in the kernel approximation. Moreover, depending on the kernel chosen and the parameter tuning for this kernel, the algorithm might fail to compute the right distance depending on the distributions at hand. Extensions have been proposed to mitigate these issues such as adding a parabola to the

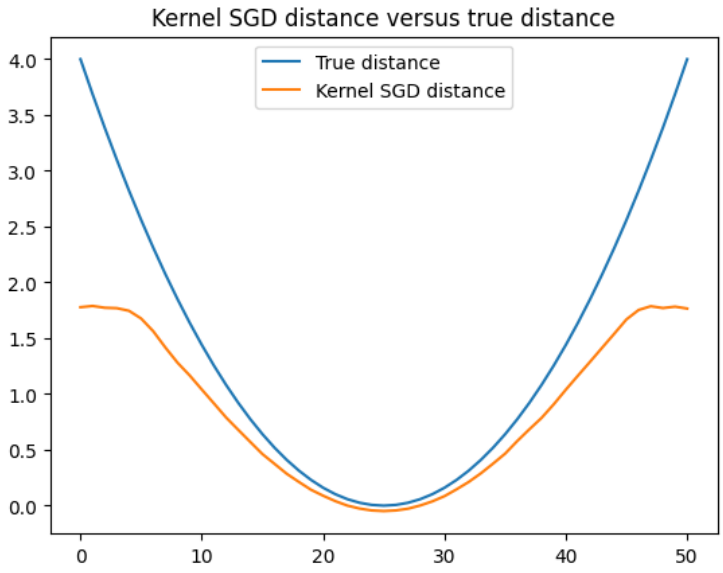

Figure 9: Continuous experiment result.

approximation $f(x) = ax^2 + bx + c + \sum \alpha_k k(x, x_k)$ of directly using a neural network to approximate f and g.

As for the semi-discrete case, we defend the argument that they are few applications of continuous optimal transport. With the limitations exposed in mind, we argue that sampling from continuous distributions and using large scale discrete solver is still a better option, at least for data science applications.

5 Conclusion

We argue that the studied paper marks an important step towards large scale solvers for Optimal Transport. As discussed earlier, we don't consider the semi-discrete and continuous cases as very important in practice, but they might provide an interesting starting point for domain specific applications where a discrete approximation is not an option. In data sciences however, the discrete case is ubiquitous. The challenge consists in the very large size and dimension of datasets. In this case, the contribution of the paper is of great importance. Correctly tuned, the SAG proposed has been shown empirically to improve the convergence speed of the Sinkhorn algorithm. This is an important step towards democratizing the use of Optimal Transport in

data sciences.

On a more critical note, we would like to stress a few limitations of optimal transport that are not discussed in the original paper. First, the entropic regularization that we often use in practice introduces a bias in the solution of Optimal Transport. The effect of this bias, especially in high dimension, should be more thoroughly addressed in the literature. More generally, using OT to fit distributions, as is often the objective in machine learning, comes with strong limitations. Indeed, as discussed in [11] (Section 3.3) Optimal Transport is blind to the topology of the distributions, which can be very problematic for geometric applications. In GAN training, or any application that consists in fitting densities in a high dimensional latent space, this lack of connexity preservation is problematic. This issue is well illustrated even in small dimensions. See Figure 10, where the model distribution α is first "split" then re-constructed over the dataset distribution β. The fitting ignores the original topology of the distributions, even if there is a natural way to match them together.

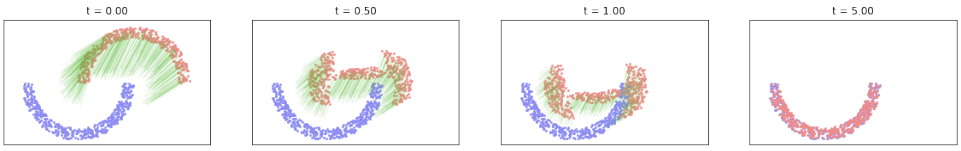

Figure 10: Gradient flow of a density fitting using OT cost between densities α (in red) and β (in blue).

6 Connection with the course

The continuous Kantorovich formulation of Optimal Transport (1) and it's dual (3) are presented in the course notes (Section 3.2, Section 5.2). The entropic regularization was discussed in Section 4.4, and was the subject of a numerical tour. The use of c-transforms to derive the semi dual formulations (5, 6) was also briefly discussed at the end of the course, see Section 6. Semi-discrete Optimal Transport was the subject of an optional numerical tour. The course notes and the book provided all the necessary tools to understand the expectation maximization formulation (7, 8) and thus, the algorithms introduced in the rest of the paper.

Reproducibility

You can reproduce all our figures using the notebook publicly available at : `https://github.com/eugeneberta/MVA/blob/main/S1_Optimal_Transport/Project/project.ipynb`

Acknowledgments

We would like to thank the authors of the studied paper for their interesting work.

References

[1] Genevay Aude, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport, 2016.

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[3] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 2017.

[4] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances. 2013.

[5] Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Annals of Mathematical Statistics*, 35:876–879, 1964.

[6] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences, 2018.

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

[8] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient, 2013.

[9] Gabriel Peyré. Course notes on computational optimal transport, 2021.

[10] Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

[11] Jean Feydy. Geometric data analysis, beyond convolutions, 2020.