Stochastic optimization for large scale optimal transport

Project report

Hugo Cisneros

January 7, 2019

This report will study the matter of applying stochastic optimization techniques to solve optimal transport problems in the discrete and semi-discrete settings.

In the discrete setting, the standard solver of the regularized OT problem is the Sinkhorn-Knopp algorithm which has a general computational complexity of $O(n^2)$. The problem is notoriously hard to solve, and the complexity of the Sinkhorn-Knopp algorithm is too high for a very large scale setting. Stochastic algorithms can be used to cope with that limitation and compute solutions of the regularized OT problem with a computational complexity of $O(n)$.

In the semi-discrete setting, some solvers exist but can only be applied to specific subproblems in low dimension and with simple cost functions. The stochastic gradient algorithm can be used for finding the optimal solution of such problems in the general form.

To explore the properties of stochastic optimization for OT, we will present three experiments: a benchmark of stochastic algorithms on synthetic data and an image retreival task that will provide some numerical results to illustrate the use of gradient aggregation algorithms (namely SAG and SAGA) for solving dicrete optimal transport problems, and a socio-economic data analysis that will provide insights into the scope of stochastic gradient algorithms for solving semi-discrete OT problems.

Contents

1 Introduction

Optimal Transport (OT) is well known for its many applications in various domains, especially when working in spaces of probability distributions. It has recently had major successes when applied to several machine learning problems in Computer Vision [1, 18] or Natural Language Processing [12]. Although efficient ways of solving OT problems exist and are widely used in practice, they are impractical for very large scale settings, which creates the need for more efficient methods. Stochastic optimization is also an essential tool at the basis of numerous successes of machine learning and its immense developments. They allow to solve very large scale problem with reasonable time and memory requirements, which make them ideal for cases where traditional methods for solving OT problems fail.

1.1 Optimal transport : problem formulations

Optimal transport dates back to 1781, when Monge studied the mathematical properties of problems involving the displacement of earth [13]. The problem has been first formulated in its modern form by Kantorovitch [11], allowing for continuous displacement of mass. OT can be interpreted as a way of defining a metric among probability distributions, called the *Wasserstein* of *earth mover's* distance. It is often described as a natural way to leverage the geometry of a space and define a metric on probability distributions, by opposition to the Euclidean distance and Kullback-Leibler divergence.

1.1.1 Entropic regularization of OT

We consider two measures $\mu \in \mathcal{M}_+^1(\mathcal{X})$ and $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ defined on metric spaces $\mathcal{X}$ and $\mathcal{Y}$. The cost of moving a unit of mass from $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ is defined by the continuous function $c \in \mathcal{C}(\mathcal{X}, \mathcal{Y})$, and written $c(x, y)$. We also define the set of joint probability measures on $\mathcal{X} \times \mathcal{Y}$

$$\Pi(\mu, \nu) \triangleq \{\pi \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y}); \forall (A, B) \subset \mathcal{X} \times \mathcal{Y}, \pi(A, \mathcal{Y}) = \mu(A), \pi(\mathcal{X}, B) = \nu(B)\}$$

The entropic regularized version of the OT problem [4] can be written as a single convex optimization problem in the following form: $\forall (\mu, \nu) \in \mathcal{M}_+^1(\mathcal{X}) \times \mathcal{M}_+^1(\mathcal{Y})$,

$$W_\varepsilon(\mu, \nu) \triangleq \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X},\mathcal{Y}} c(x, y) \mathrm{d}\pi(x, y) + \varepsilon \mathrm{KL}(\pi || \mu \otimes \nu) \qquad (\mathcal{P}_\varepsilon)$$

With $\mathrm{KL}(\pi||\mu \otimes \nu)$ corresponding to the Kullback-Leibler divergence between measures π and $\mu \otimes \nu$, defined by $\mathrm{KL}(\pi||\xi) \triangleq \int_{\mathcal{X},\mathcal{Y}} \left(\log(\frac{d\pi}{d\xi}(x, y) - 1 \right) d\xi(x, y)$.

For $\varepsilon > 0$, the above problem is strongly convex. $(\mathcal{P}_\varepsilon)$ is usually called the primal form of the regularized OT problem, by opposition to the dual and semi-dual form that will be studied further.

Sinkhorn for discrete OT In the discrete setting $\mu = \sum_i^n \delta_{x_i} \boldsymbol{\mu}_i$ and $\nu = \sum_j^m \delta_{x_j} \boldsymbol{\nu}_j$, the sums are finite and the cost is $\mathbf{C} \in \mathbb{R}^{n \times m}$. The structure of the KL divergence gives the optimal solution $\mathbf{P}_\varepsilon \in \Pi(\mu, \nu)$ a convenient structure that makes it possible solving the problem using Sinkhorn's algorithm [4]. There indeed exist two scaling variables $\mathbf{u}_\varepsilon \in \mathbb{R}^n$ and $\mathbf{v}_\varepsilon \in \mathbb{R}^m$ such that

$$\mathbf{P}_\varepsilon = \mathrm{diag}(\mathbf{u}_\varepsilon) \mathbf{K}_\varepsilon \mathrm{diag}(\mathbf{v}_\varepsilon)$$

Where $(\mathbf{K}_\varepsilon)_{i,j} = \exp(-\mathbf{C}_{i,j}/\varepsilon)$ [16]. Those scaling variables can be computed iteratively with the following update at step ℓ,

$$\mathbf{u}_\varepsilon^{\ell+1} \triangleq \frac{\mu}{\mathbf{K}_\varepsilon \mathbf{v}_\varepsilon^\ell} \quad \text{and} \quad \mathbf{v}_\varepsilon^{\ell+1} \triangleq \frac{\nu}{\mathbf{K}_\varepsilon^T \mathbf{u}_\varepsilon^{\ell+1}} \tag{1}$$

Because each step of the algorithm relies on a vector-matrix computation, the overall complexity of the algorithm is $O(nm)$ in the most general configuration. Moreover, the rate of convergence of the algorithm is linear in the number of iterations [6].

Algorithm 1 Sinkhorn algorithm

1: Data: $\mathbf{C}$, ε, μ, ν
2: $\mathbf{K} \leftarrow \exp(-\mathbf{C}/\varepsilon)$
3: $\mathbf{u} = \mathbf{v} = 0$
4: **while** !(`stopping_criterion`) **do**
5: $\quad \mathbf{u} \leftarrow \frac{\mu}{K\mathbf{v}}$
6: $\quad \mathbf{v} \leftarrow \frac{\nu}{K^T\mathbf{u}}$
7: **end while**
8: **return** $\mathrm{diag}(\mathbf{u})\mathbf{K}\mathrm{diag}(\mathbf{v})$

The algorithm can be used in a large scale setting by making use of specific hardware (multiple Wasserstein distances can be computed in parallel on a GPU [22]) and in some other specific cases where the kernel $\mathbf{K}$ is separable or can be expressed as a convolution [16]. In the general case however, the computational complexity of Sinkhorn's algorithm can be prohibitively large for large scale problems.

The solution of an OT problem in the discrete setting can be represented as a transportation matrix $\mathbf{P}_\varepsilon$. An example of such a solution can be visualized along with two randomly generated measures on $1, ..., 50$ on Figure 1. The distance matrix is the L_2 distance on the set $1, ..., 50$, the value of ε was set to 1e-4 and the problem was solved with Sinkhorn's algorithm.

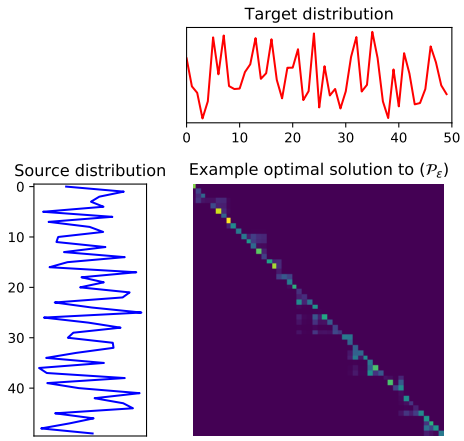


Figure 1: An example solution to discrete regularized OT for the *blue* and *red* discrete measures.

Naturally, most of the transport is concentrated close to the diagonal of the matrix, that is where the cost is smallest in the example.

1.1.2 Dual and Semi-dual formulations

The dual and semi-dual variations of the OT problem are essential for constructing and applying stochastic optimization methods to solve it.

We define the following set for $c \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})$

$$U_c \triangleq \{(u, v) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{X}); \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, u(x) + v(y) \leq c(x, y)\}$$

The dual problem can be derived using Fencher-Rockafellar's theorem [8] and reads

$$W_\varepsilon(\mu, \nu) = \max_{(u,v) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} u(x) \mathrm{d}\mu(x) + \int_{\mathcal{Y}} v(y) \mathrm{d}\nu(y) - \iota^\varepsilon_{U_c}(u, v) \qquad (\mathcal{D}_\varepsilon)$$

Where we have defined $\iota^0_{U_c} = \iota_{U_c}$ and for $\varepsilon > 0$, $\iota^\varepsilon_{U_c}$ is the *smoothed* approximation of the indicator function of U_c,

$$\iota^\varepsilon_{U_c}(u, v) = \varepsilon \int_{\mathcal{X} \times \mathcal{Y}} \exp\left(\frac{u(x) + v(y) - c(x, y)}{\varepsilon}\right) \mathrm{d}\mu(x) \mathrm{d}\nu(y)$$

If we write the optimality conditions in v of $(\mathcal{D}_\varepsilon)$, we get the following relation between u and v

$$\forall x \in \mathcal{X}, u(x) = v^{c,\varepsilon}(x) \triangleq \begin{cases} \min_{y \in \mathcal{Y}} c(x, y) - v(y) & \text{if } \varepsilon = 0 \\ -\varepsilon \log\left(\int_{\mathcal{Y}} \exp\left(\frac{v(y) - c(x, y)}{\varepsilon}\right)\right) & \text{if } \varepsilon > 0 \end{cases}$$

$v^{c,\varepsilon}$ is sometimes called the c-transform [3] of dual variable v. By plugging this expression back into $(\mathcal{D}_\varepsilon)$, we get the semi-dual form of the OT problem

$$W_\varepsilon(\mu, \nu) = \max_{v \in \mathcal{C}(\mathcal{Y})} H_\varepsilon(v) \triangleq \int_{\mathcal{X}} v^{c,\varepsilon}(x) \mathrm{d}\mu(x) + \int_{\mathcal{Y}} v(y) \mathrm{d}\nu(y) - \varepsilon \qquad (\mathcal{S}_\varepsilon)$$

This formulation is crucial to solving the semi-discrete OT problem, because optimization can be done with respect to the discrete dual variable instead of the continuous one.

Dual variables and Sinkhorn algorithm In the discrete setting, the scaling variables $\mathbf{u}$ and $\mathbf{v}$ of the Sinkhorn algorithm can be linked to the dual variables u and v of $(\mathcal{D}_\varepsilon)$ with the relation

$$(\mathbf{u}, \mathbf{v}) = (\exp(u/\varepsilon), \exp(v/\varepsilon))$$

The proof of this result can be found in [16]. With these variables, the Sinkhorn algorithm can be re-written as a *block coordinate ascent* strategy on the dual variables.

We also note that, since the scaling variables of the Sinkhorn algorithm are defined up to a multiplicative constant $\lambda > 0$, the dual variables are also defined up to an additive constant.

Furthermore, for a v solving $(\mathcal{S}_\varepsilon)$, an optimal u can be recovered with $u = v^{c,\varepsilon}$ (this also shows that u and v can be translated by an arbitrary constant in opposite directions). From a pair (u, v) solving $(\mathcal{D}_\varepsilon)$, an optimal solution π of $(\mathcal{P}_\varepsilon)$ can be computed with $\mathrm{d}\pi(x, y) = \exp\left(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}\right) \mathrm{d}\mu(x) \mathrm{d}\nu(y)$.

1.2 Relevant Work

The unregularized Kantorovitch formulation [11] of optimal transport is usually solved as a large scale linear program in the case of finite distributions (for example a network simplex algorithm). Some heuristics were introduced to cope with the relatively high computational cost of those method, such as pruning the long distances between histogram bins [15] or using quad-trees and random shifts such as in [21].

Since the regularized version of the problem is strongly convex, it can benefit from all the optimization methods associated with and advantages of this property. As presented above, regularized optimal transport problems are usually tackled with Sinkhorn's algorithm [4]. Its linear convergence rate and paralellisation properties (showcased in [23]) make it the prevalent way of computing solutions to the problem.

For semi-discrete OT, the main way to compute solution to the problem is by using semi-discrete solvers such as [2].

The implementations and stochastic formulations of discrete and semi-discrete optimal transport problems are based on Genevay et al.'s paper on stochastic optimization applied to optimal transport [8]. This paper also explores the continuous-continuous setting, which has comparatively to the other two been less explored in practice because it involves estimating functions, which is a harder task in general. [8] uses a Kernel expansion of the dual variables to solve the continuous-continuous problem, while [1] and [20] have used neural networks to represent the dual variables.

1.3 Contributions

This contributions of this paper are the following:

- Lay out the stochastic formulations of discrete and semi-discrete OT problems proposed in [8].
- Implement and benchmark the SAG (Stochastic Average Gradient) algorithm [19] in the discrete setting on synthetic and real data, give some supplementary results to [8].
- Implement and benchmark the SAGA algorithm [5] in the same setting and compare its performances with SAG.
- Implement and benchmark the SGD algorithm for semi-discrete optimal transport on a socio-economic case study.

We show that stochastic optimization is a very useful tool for dealing with OT problems, as it provides a possibly more efficient way to compute solutions in a discrete setting and a general way of solving semi-discrete OT problems.

2 A numerical study of stochastic optimization for large scale optimal transport

As presented in the last section, Sinkhorn's algorithm has an relatively high computational complexity that can prohibit its utilization in a very large scale setting. With their linear complexity in general, stochastic optimization algorithms offer an attractive alternative to Sinkhorn, provided the problem can be framed as optimization of an empirical or expected risk. In this section, we will quickly present this formulation, and then study and benchmark the algorithms presented for discrete and semi-discrete optimal transport. All the code used for generating the results of this section is available online[1].

2.1 Stochastic optimization for OT

2.1.1 Stochastic formulation

To make it possible using stochastic optimization methods for the OT problems, we frame the dual ($\mathcal{D}_\varepsilon$) and semi-dual ($\mathcal{S}_\varepsilon$) optimization problems as maximization of expectations,

$$W_\varepsilon(\mu,\nu) = \max_{(u,v)} \ \mathbb{E}_{X,Y}[f_\varepsilon(X,Y,u,v)] \qquad (\mathbb{E}\mathcal{D}_\varepsilon)$$

$$= \max_{v} \ \mathbb{E}_{X}[h_\varepsilon(X,v)] \qquad (\mathbb{E}\mathcal{S}_\varepsilon)$$

Where the two independent random variables X and Y are defined on $\mathcal{X}$, $\mathcal{Y}$, follow respectively the distributions μ and ν, and with the following definitions for f_ε and h_ε,

$$\forall \varepsilon > 0, \ f_\varepsilon(x,y,u,v) = u(x) + v(y) - \varepsilon \exp\left(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}\right)$$

$$\forall \varepsilon \geq 0, \ h_\varepsilon(x,v) = \int_{\mathcal{Y}} v(y)\mathrm{d}\nu(y) + v^{c,\varepsilon}(y) - \varepsilon$$

2.1.2 Stochastic optimization algorithms

Three algorithms will be studied for solving the stochastic OT problems. The first is the standard Stochastic Gradient algorithm (SG). It is based on sampling a realization of the random variable in the expected risk (X on our case) and use it as an estimate of the gradient direction for the update of the iterate.

In the case of problem ($\mathbb{E}\mathcal{S}_\varepsilon$), Algorithm 2 shows the details of each step. The + symbol on line 6 corresponds to a gradient ascent since the objective has to be maximized.

The two other algorithms studied in the report, Stochastic Averaged Gradient (SAG) [19] and SAGA [5] belong to the family of gradient aggregation algorithms. These were designed for finite sum objectives (empirical risks), which is only applicable in our case to the discrete optimization problem as we will see further.

The two methods are based on the idea that keeping a history of past gradient in a stochastic context could give better estimate of the real gradient and thus ensure faster convergence. This intuition is verified, because both SAG and SAGA guarantee a $O(1/k)$ convergence rate in the general case and a linear convergence rate in the strongly convex case [19, 5](against $O(1/\sqrt{k})$ and $O(1/k)$ respectively for

[1] On GitHub, at the URL `https://github.com/hugcis/large-scale-optimal-transport`

Algorithm 2 SG algorithm

Choose initial iterate v_0
for $k = 0, ..., K$ **do**
    Generate a realization x_k of X
    Compute the stochastic vector $\nabla h_\varepsilon(x_k, v_k)$
    Choose step size α_k
    $v_{k+1} \leftarrow v_k + \alpha_k \nabla h_\varepsilon(x_k, v_k)$
end for

standard SG for a suitable decreasing step size sequence and standard assumptions [14]), with automatic adaptation to the local strong-convexity of the objective.

Those properties make them very attractive for applications where SG shows its limitation, and [8] has presented the advantages of using SAG for discrete optimal transport. One major drawbacks of the gradient aggregation algorithms is the fact that all gradients have to be stored at each step of the algorithm, which can be a strong limitation when working with very large scale problems.

2.2 Discrete Optimal Transport

In the case of discrete OT, we recall that μ and ν can both be written as finite sums of Diracs, i.e $\mu = \sum_{i=1}^{I} \mu_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{J} \nu_j \delta_{y_j}$ with $\forall i, x_i \in \mathcal{X}; \forall j, x_j \in \mathcal{Y}$ and $\mu \in \Sigma_I, \nu \in \Sigma_J$.

2.2.1 Details of formulation

We re-write the primal, dual and semi-dual problems with discrete measures for $\varepsilon > 0$:

$$W_\varepsilon(\mu,\nu) = \min_{\boldsymbol{\pi} \in \boldsymbol{U}_{\mu,\nu}} \sum_{i,j=1}^{I,J} \mathbf{C}_{i,j}\boldsymbol{\pi}_{i,j} + \varepsilon \sum_{i,j} \left(\log \frac{\boldsymbol{\pi}_{i,j}}{\boldsymbol{\mu}_i \boldsymbol{\nu}_j} - 1 \right) \boldsymbol{\pi}_{i,j} \tag{$\overline{\mathcal{P}}_\varepsilon$}$$

$$= \max_{(\boldsymbol{u},\boldsymbol{v}) \in \mathbb{R}^I \times \mathbb{R}^J} \sum_{i=1}^{I} \boldsymbol{u}_i \boldsymbol{\mu}_i + \sum_{j=1}^{J} \boldsymbol{v}_j \boldsymbol{\nu}_j - \varepsilon \sum_{i,j} \exp\left(\frac{\boldsymbol{u}_i + \boldsymbol{v}_j - \boldsymbol{C}_{i,j}}{\varepsilon} \right) \boldsymbol{\mu}_i \boldsymbol{\nu}_j \tag{$\overline{\mathcal{D}}_\varepsilon$}$$

$$= \max_{\boldsymbol{v} \in \mathbb{R}^J} \overline{H}_\varepsilon(\boldsymbol{v}) = \sum_{i=1}^{I} \overline{h}_\varepsilon(x_i, \boldsymbol{v}) \boldsymbol{\mu}_i \tag{$\overline{\mathcal{S}}_\varepsilon$}$$

Where we have defined

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

Equation ($\overline{\mathcal{S}}_\varepsilon$) can be interpreted as an empirical risk associated with the functions ($\overline{h}_\varepsilon(x_i, \cdot)$). We can therefore apply stochastic optimization algorithms on the problem ($\overline{\mathcal{S}}_\varepsilon$). The gradient to be computed is

$$\nabla_{\boldsymbol{v}} \overline{h}_\varepsilon(x, \boldsymbol{v})_j = \boldsymbol{\nu}_j - \exp\left(\frac{\boldsymbol{v}_j - c(x, y_j)}{\varepsilon} \right) \left(\sum_{i=1}^{J} \exp\left(\frac{\boldsymbol{v}_i - c(x, y_i)}{\varepsilon} \right) \right)^{-1}$$

2.2.2 Stochastic algorithms and discrete OT

As explained in Section 2.1.2, SAG and SAGA have comparatively better convergence guarantees than SG for minimizing or maximizing strongly convex and non-strongly convex objectives.

Algorithms 3 and 4 show the pseudo-code implementation of both algorithms in the case of discrete OT. The parameter `step` designates a learning rate that is usually chosen in accordance with the Lipschitz constant of the function $\overline{H}_\varepsilon$ to be maximized, $L = \max_i \boldsymbol{\mu}_i/\varepsilon$.

Algorithm 3 SAG algorithm for discrete OT

Choose initial iterate $\boldsymbol{v}_1$
$\mathbf{avg} \leftarrow 0_J, \mathbf{grad}_i \leftarrow 0_J$
(Note: Gradient can be initialized with a pass of SG)
for $k = 1, ..., K$ **do**
    Sample uniformly $i \in \{1, ..., I\}$
    $\mathbf{avg} \leftarrow \mathbf{avg} - \frac{1}{I}\mathbf{grad}_i$
    $\mathbf{grad}_i \leftarrow \boldsymbol{\mu}_i \nabla_v \overline{h}_\varepsilon(x, \boldsymbol{v}_k)$
    $\mathbf{avg} \leftarrow \mathbf{avg} + \frac{1}{I}\mathbf{grad}_i$
    $\boldsymbol{v}_{k+1} \leftarrow \boldsymbol{v}_k + \mathtt{step} * \mathbf{avg}$
end for

Algorithm 4 SAGA algorithm for discrete OT

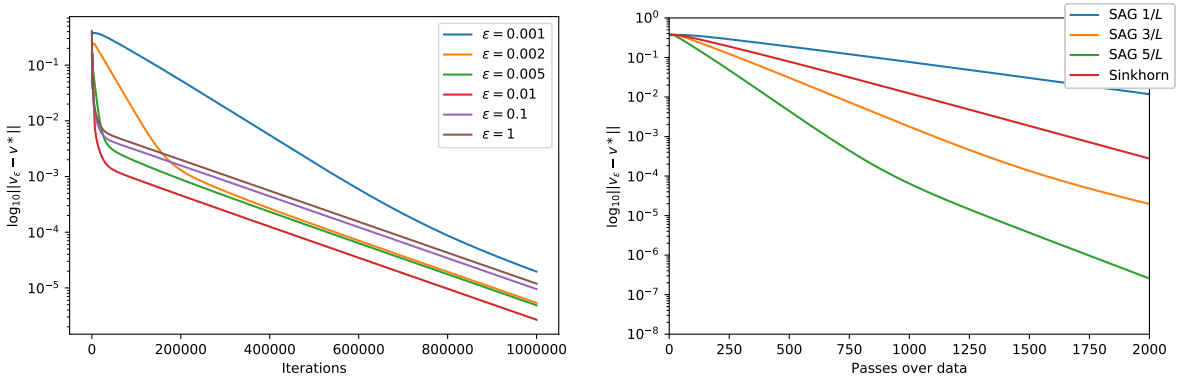
Choose initial iterate $\boldsymbol{v}_1$
$\mathbf{avg} \leftarrow 0_J, \mathbf{grad}_i \leftarrow 0_J, \mathbf{last\_grad} \leftarrow 0_J$
(Note: Gradient can be initialized with a pass of SG)
for $k = 1, ..., K$ **do**
    Sample uniformly $i \in \{1, ..., I\}$
    $\mathbf{last\_grad} \leftarrow \mathbf{grad}_i$
    $\mathbf{avg} \leftarrow \mathbf{avg} - \frac{1}{I}\mathbf{last\_grad}$
    $\mathbf{grad}_i \leftarrow \boldsymbol{\mu}_i \nabla_v \overline{h}_\varepsilon(x, \boldsymbol{v}_k)$
    $\mathbf{avg} \leftarrow \mathbf{avg} + \frac{1}{I}\mathbf{grad}_i$
    $\boldsymbol{v}_{k+1} \leftarrow \boldsymbol{v}_k + \mathtt{step} * (\mathbf{grad}_i - \mathbf{last\_grad} + \mathbf{avg})$
end for

We first evaluate those algorithms on the toy problem of computing the solution to $(\mathcal{P}_\varepsilon)$ for two random discrete measures with support $\{1, ..., N\}$. They are obtained by sampling N values from a uniform distribution on $[0, 1]$ twice and normalizing the histograms in order to obtain measures.

Observed convergence is linear after a certain number of passes over the data in those examples, which confirm the theoretical guarantees of the SAG and SAGA algorithm.

As recommended in [19], the three learning rates $1/L$, $3/L$ and $5/L$ are tested for the SAG algorithm. Depending on the chosen learning rate, improvement over Sinkhorn's algorithm can be obtained or not, which is a crucial information when trying to apply the algorithms on very large scale settings. Notably here, both $3/L$ and $5/L$ show very significant improvements over Sinkhorn.

For the SAGA algorithm, the theoretical convergence guarantee is proven for a learning rate of $1/3L$. We try this learning rate along with some others such as



(a) Convergence of SAG algorithm for different values of ε.

(b) Convergence of SAG algorithm for different learning rate and $\varepsilon = 0.001$.

Figure 2: Convergence of SAG on two randomly generated histograms of size 500. 50 independent runs were averaged.

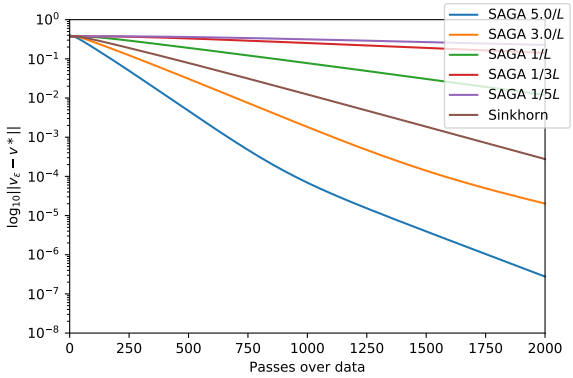


Figure 3: Convergence of SAGA algorithm for different learning rates on two randomly generated histograms of size 500. 50 independent runs were averaged.

$1/5L$, $1/L$, $3/L$ and $5/L$. We note that, although no convergence guarantees exist for these values, SAGA algorithm has a behavior very similar to the SAG algorithm for the same pairs of learning rate $(1/nL)$. This implies that SAGA doesn't show any significant advantage or disadvantage over SAG for this particular example.

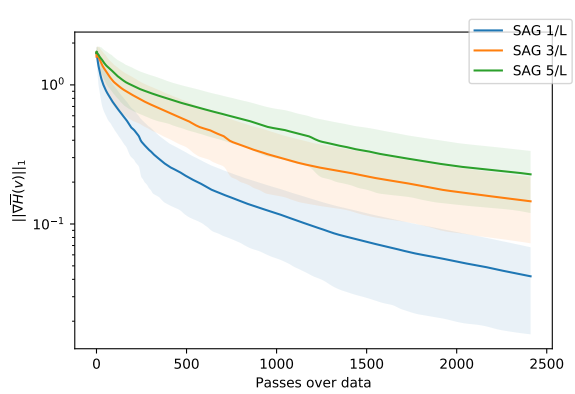
The optimal values all convergence rate are compared against were obtained by running Sinkhorn's algorithm up to convergence of the solution to machine precision.

We now study the properties of stochastic optimization on a real world dataset, for the task of image retrieval described in [18]. For this, we work with the INRIA Holiday dataset [10][2]. The pictures are preprocessed and converted from RGB to the CIE-Lab color space [24], which was in turn uniformly quantized into 32^3 bins in a 3D histogram. Since much of those bins are empty (the range of color of a dataset of natural images is limited and RGB doesn't capture all the color space available from the CIE-Lab color space), we prune the empty bins in the histograms and represent all images with histograms of size 4128. We note that by opposition to the last experiment, representation of the images in this form are typically very sparse. The cost function is the L_2 norm over the CIE-Lab color space.

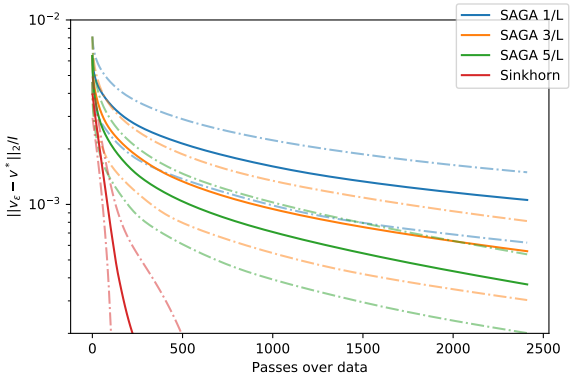
With this representation, we compute 10 pairwise distances between images from

[2]The dataset is available for download at `http://lear.inrialpes.fr/~jegou/data.php#holidays`.

the dataset selected at random and show the convergence results for SAG, SAGA and Sinkhorn. ε was set to 0.01. We compare the results to an optimal dual variable $\boldsymbol{v}^*$ which corresponds to the best obtained dual variable across all methods.

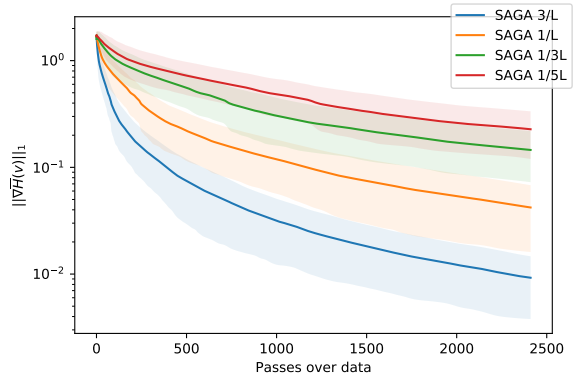


(a) Evolution of $||\nabla\overline{H}(v)||_1$.

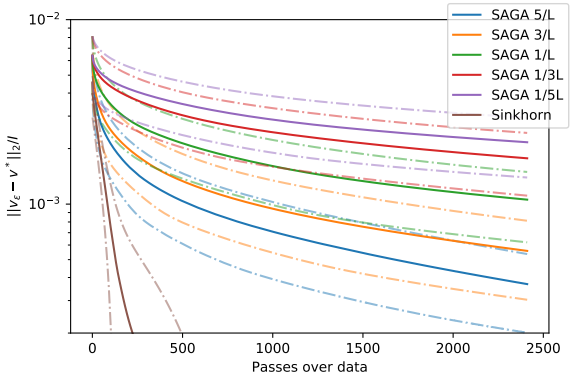


(b) Convergence of SAG algorithm for different learning rates.

Figure 4: Convergence of SAG on 10 pairs of images. Dashed lines and filled areas represent deviation from the mean (solid line).



(a) Evolution of $||\nabla\overline{H}(v)||_1$.



(b) Convergence of SAG algorithm for different learning rates.

Figure 5: Convergence of SAGA on 10 pairs of images. Dashed lines and filled areas represent deviation from the mean (solid line).

In this example, Sinkhorn was consistently faster than the other optimization methods which had a convergence rate close to $O\left(\frac{1}{k}\right)$. For SAG and SAGA, the overall best performing learning rate was $5/L$ in terms of speed of convergence. It however gave some poor results for SAGA in some cases where the gradient did not converge to 0.

2.3 Semi-discrete Optimal Transport

In semi-discrete optimal transport, μ can be an arbitrary measure and the other measure $\nu = \sum_{j=1}^{J} \nu_j \delta_{y_j}$ is discrete [16]. We therefore need to work with the expectation form of the dual and semi dual problem presented in 2.1.1. We recall here

the semi-dual form of the problem ($\mathbb{E}\mathcal{S}_\varepsilon$):

$$W_\varepsilon(\mu, \nu) = \max_v \ \mathbb{E}_X[h_\varepsilon(X, v)]$$

X follows the law of μ in that case, and h_ε has been defined above. The gradient aggregation algorithms such as SAG and SAGA cannot be applied as is to such problem, but the SG algorithms still fits to its constraints, since it only needs a random variable to sample from.

A way to fall back to a setting were the gradient aggregation family of algorithms is applicable is to discretize the continuous measure and approximate it with a discrete one. This method has the drawback of making the obtained solution inexact due to discretization noise, and is therefore not studied in this report.

2.3.1 Stochastic gradient algorithm for semi-discrete OT

The SG algorithm applied to the semi-discrete setting is detailed in algorithm 5. The convergence rate is guaranteed to be $O(1/\sqrt{k})$ by using averaging of the iterates $\boldsymbol{v}$ [17] (line 5 in algorithm 5) since the problem is not strongly convex.

Algorithm 5 SG algorithm for semi-discrete OT

1: $\tilde{\boldsymbol{v}} \leftarrow 0_J, \boldsymbol{v} \leftarrow \tilde{\boldsymbol{v}}$
2: **for** $k = 1, ..., K$ **do**
3: Sample x_k realization of X following μ
4: $\tilde{\boldsymbol{v}} \leftarrow +\frac{\text{step}}{\sqrt{k}} \nabla_v \overline{h}_\varepsilon(x_k, \tilde{\boldsymbol{v}})$
5: $\boldsymbol{v} \leftarrow \frac{1}{k}\tilde{\boldsymbol{v}} + \frac{k-1}{k}\boldsymbol{v}$
6: **end for**

2.3.2 Case study of semi-discrete OT for analysis of socio-economic data

Problem formulation We now study the properties of SGD algorithm applied to a semi-discrete OT problem based on real-world data. This experiment is inspired from Hartmann and Schuhmacher 's description of the delivery resource allocation problem [9] and Galichon's book on optimal transport for economics [7].

We will study here a problem of resource allocation based on open datasets available from the official French open dataset repository `data.gouv.fr`. Similarly to the delivery resource allocation problem studied in [9], we consider a limited resource spatially scattered and a demand density across a territory. More specifically, we consider here a set of J middle schools as the ressource (that is indeed spatially distributed), that we can assimilate to a discrete measure over a territory. That measure is supported by $\mathcal{Y} = \{1, .., J\}$.

To model demand, we use the population density μ (in the sense of number of people divided by the occupied area), with the assumption that demand for school is roughly proportional to the population density of the area. The support of this continuous measure is a delimited territory. To model the transportation cost of people going to a school, we use the square distance on the territory, or L_2 norm. For any inhabitant x the cost to go to a school located at y is therefore $\Phi(x, y) = |x-y|^2$. The support of μ is written $\mathcal{X}$.

Supply capacity for each school is represented by the the total number of students at the school. The total school supply sums to one which equates the total demand.

A rudimentary way of solving this ressource allocation problem would be to only take into account the supply and demand without modelling any form of side-effect

of high demand (such as high price, or in our case, the high entry level of the school). Everybody would then choose the school such as to maximize a *utility* function $u(x) = \max_{j\in\{1,\dots,J\}} -\Phi(x, y)$. The set of people preferring school j over other schools is

$$\mathcal{X}_j^0 \triangleq \{x \in \mathcal{X} \mid \Phi(x, y_j) \leq \Phi(x, y_k), \forall k\}$$

This corresponds to a Voronoi tessellation of the territory with the schools being the centers of the Voronoi cells. Demand for school j is $\mathbb{P}(x \in \mathcal{X}_j) = \int_{\mathcal{X}_j} \mathrm{d}\mu(x)$.

Now, if we consider that school have an entry level $\mathbf{p} = (\mathbf{p}_j)_{j\in\{1,\dots,M\}}$ that changes with demand, and we assume that utility is written[3]

$$u(x) = \max_{j\in\{1,\dots,J\}} -\mathbf{p}_j - \Phi(x, y_j)$$

We write q_j the demand for school j. At market equilibrium, demand for a school equates the supply it provides and all schools are at complete capacity, hence $\mathbb{P}(x \in \mathcal{X}_j) = q_j$, with $\sum_j q_j = 1$.

A transportation problem A planner seeks to find an optimal assignment π that minimizes the total transportation cost, while matching the population density with the corresponding schools. She would have

$$\min_{\pi\in\Pi(\mu,q)} \int_{\mathcal{X},\mathcal{Y}} c(x, y)\mathrm{d}\pi(x, y)$$

This is the exact formulation of a semi-discrete transport problem with continuous measure μ and discrete measure $q = \sum_j \delta_{y_j} q_j$.

We now write the semi-dual form of the entropic regularization of the problem

$$W_\varepsilon(\mu, q) = \max_{v\in\mathbb{R}^J} \mathbb{E}_X[\overline{h}_\varepsilon(X, \boldsymbol{v})]$$

We recall the expression of $\overline{h}_\varepsilon$ and give some details about it in the setting of the case study.

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

Note also that the solution to the optimization problem is a smooth version of the Laguerre cells $\mathbb{L}_{\boldsymbol{v}}(y_j) \triangleq \{x \in \mathcal{X} \mid c(x, y_j) - v_j \leq c(x, y_k) - v_k, \forall k\}$. It seems now that the semi-dual problem can be interpreted in terms of choosing a entry level $\mathbf{p}$ for all schools, by identifying $\mathbf{p}$ with $-\boldsymbol{v}$. $\overline{h}_\varepsilon$ then becomes

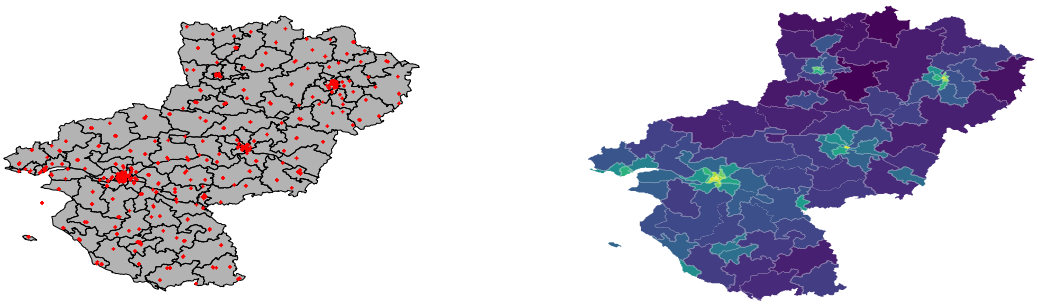
$$\overline{h}_\varepsilon(x, \mathbf{p}) = -\sum_{j=1}^{J} \mathbf{p}_j q_j + \begin{cases} -\varepsilon \log\left(\sum_{j=1}^{J} \exp\left(\frac{-\mathbf{p}_j - c(x,y_j)}{\varepsilon}\right) q_j\right) - \varepsilon & \text{if } \varepsilon > 0, \\ \min_j(c(x, y_j) + \mathbf{p}_j) & \text{if } \varepsilon = 0 \end{cases}$$

The problem therefore amounts to finding the price that maximizes the expected utility $u(x) = \min_{j\in\{1,\dots,J\}} \mathbf{p}_j + \Phi(x, y_j)$ (and smoothed utility) of everyone while minimizing the global output of all schools (demand times the level of each school).

Possible applications of this kind of model range from defining the spatial assignment of public schools to identifying the weak points in the geographical repartition of schools and managing the attractiveness of all schools.

[3]This modelling could obviously be discussed, since the assumption that people will preferably choose a school with low entry level and closer to them is debatable. The assumption is made here to show a possible way of posing the problem in terms of optimal transport.

Numerical experiments We implement the SG algorithm to solve the semi-discrete OT problem for the two measures represented on Figure 9. We use the squared distance as our cost function. This is based on the assumption that straight line distance is a good proxy for the travel distance, which might not be very accurate when there is a body of water between two points for example. A better distance function would use isochrone curves, but we assume that the straight line distance is a good enough approximation for this example.



(a) School locations

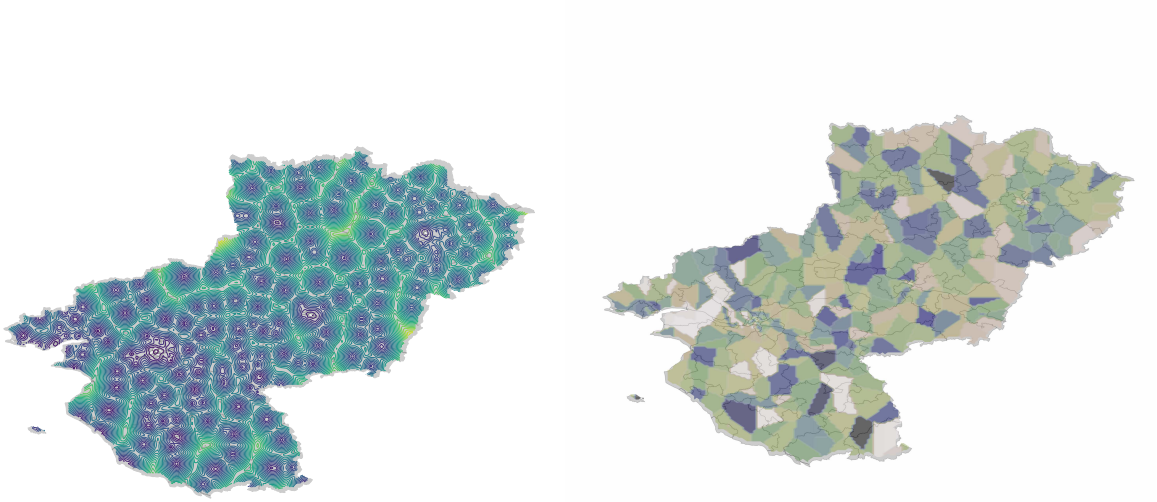
(b) Population log-density by "cantons"

Figure 6: Example initial setting of the problem described in paragraph 2.3.2 for the French administrative region "Pays de la Loire"

Setting the vector $\boldsymbol{v}$ to 0 would yield an assignment that coincides with the smoothed Voronoi tessellation of the territory. The smoothed $\bar{c}$-transform of $\boldsymbol{v}$ represents the expected utility at position x. The optimal function, computed after 10^8 iterations of SG, is represented on Figure 7a. The Laguerre cells corresponding to the computed dual potential $\boldsymbol{v}$ are displayed on Figure 7b.

Figure 8a shows the convergence plot for several values of ε of the SG algorithm on the example displayed Figure 9. Figure 8b shows a moving average of the gradients during processing, lighter lines represent the real values. Convergence is observed to be much slower than for the discrete setting with gradient aggregation algorithms, which is in accordance with theory.

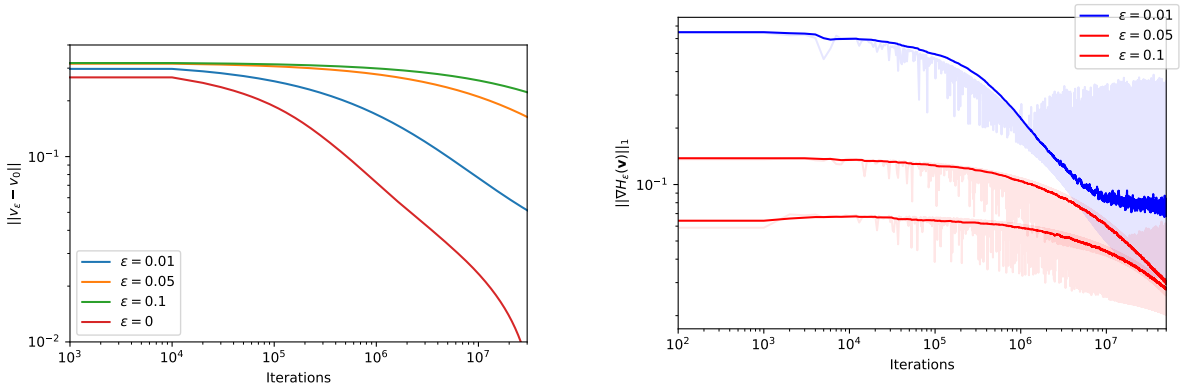
The iterates were compared against an optimal $\boldsymbol{v}_0^*$ obtained by running SG on the unregularized problem for 5×10^8 iterations (5 times more than plotted). Note that for $\varepsilon \neq 0$, the iterates are not expected to converge to the unregularized solution and will rather converge to some other value.



(a) Smoothed $\overline{c}$-transform of an optimal $\boldsymbol{v}$, $\boldsymbol{v}^{\overline{c},\varepsilon}$ for $\varepsilon = 0.01$.

(b) Laguerre cells for the computed value of $\boldsymbol{v}$, corresponding to the assigned schools that minimizes the overall transportation cost with $\varepsilon = 0.01$.

Figure 7: Results for the problem illustrated on Figure 9



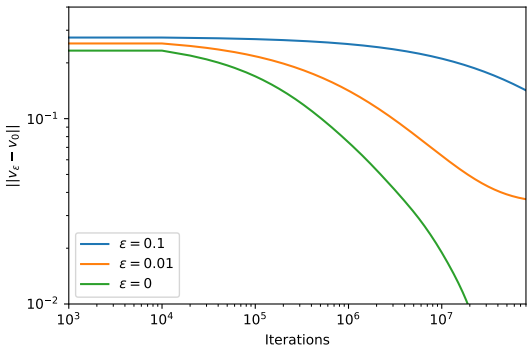
(a) Convergence towards unregularized solution.

(b) Evolution of gradient of objective function.

Figure 8: Convergence of SG in the semi-discrete setting for different values of ε

(a) Smoothed $\overline{c}$-transform of an optimal $\boldsymbol{v}$, $\boldsymbol{v}^{\overline{c},\varepsilon}$ for $\varepsilon = 0.01$.

(b) Laguerre cells.



(c) Convergence plot for several values of ε.

Figure 9: Another example of convergence for the French administrative region "Bretagne".

3 Conclusion and Perspectives

Stochastic optimization is a very promising tool for optimal transport and its application to very large scale settings.

We have presented a numerical analysis of SAG and SAGA for solving discrete OT problems on synthetic and real data for image retrieval. It has shown that these methods can sometimes outperform Sinkhorn's algorithm, but with a strong dependency on the parameters of the problems. We have also studied the stochastic gradient algorithm for semi-discrete optimal transport and presented a potential application of this method.

Both [8] and experiments in this paper show that in the discrete setting, stochastic algorithms can sometimes outperform Sinkhorn's algorithm and therefore become an essential way of solving optimal transport problems. However, although [8]'s word vectors example showed very consistent and significant improvement over Sinkhorn, it was not always observed in our experiments and seems to depend heavily on the choice of step size for the SAG and SAGA algorithms that were used to achieve this result. A more thorough exploration of the convergence properties on several other problems could be needed to have a better understanding of the influence of the size of the problem, structure of the input and output measures and algorithm's parameters.

In the semi-discrete setting, the stochastic gradient algorithm provides a general solver that can be applied to any type of cost function and is also scalable. We have presented the convergence properties of this algorithm on a real-world dataset that highlights some possible applications of semi-discrete optimal transport.

To achieve an in-depth evaluation of the stochastic methods for optimal transport, one would need to work in very large scale setups, where the $O(n^2)$ computational complexity prevents from using Sinkhorn's algorithm at all. This would allow to really evaluate how practical stochastic optimization algorithms for optimal transport are.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein GAN". In: *arXiv:1701.07875 [cs, stat]* (Jan. 26, 2017). arXiv: 1701.07875.

[2] F. Aurenhammer, F. Hoffmann, and B. Aronov. "Minkowski-Type Theorems and Least-Squares Clustering". In: *Algorithmica* 20.1 (Jan. 1998), pp. 61–76. ISSN: 0178-4617. DOI: 10.1007/PL00009187.

[3] M. Cuturi and G. Peyré. "A Smoothed Dual Approach for Variational Wasserstein Problems". In: *SIAM Journal on Imaging Sciences* 9.1 (2016), pp. 320–343. DOI: 10.1137/15M1032600. eprint: https://doi.org/10.1137/15M1032600.

[4] Marco Cuturi. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc., 2013, p. 9.

[5] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. "SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives". In: *arXiv:1407.0202 [cs, math, stat]* (July 1, 2014). arXiv: 1407.0202.

[6] Joel Franklin and Jens Lorenz. "On the scaling of multidimensional matrices". In: *Linear Algebra and its Applications* 114-115 (Mar. 1, 1989), pp. 717–735. ISSN: 0024-3795. DOI: 10.1016/0024-3795(89)90490-4.

[7] Alfred Galichon. *Optimal Transport Methods In Economics*. OCLC: 1028167041. S.l.: Princeton University Press, 2018. ISBN: 978-0-691-18346-6.

[8] Aude Genevay et al. "Stochastic Optimization for Large-scale Optimal Transport". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee et al. 2016, pp. 3440–3448.

[9] Valentin Hartmann and Dominic Schuhmacher. "Semi-discrete optimal transport - the case p=1". In: *arXiv:1706.07650 [math, stat]* (June 23, 2017). arXiv: 1706.07650.

[10] Herve Jegou, Matthijs Douze, and Cordelia Schmid. "Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search". In: *Computer Vision – ECCV 2008*. Ed. by David Forsyth, Philip Torr, and Andrew Zisserman. Vol. 5302. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 304–317. ISBN: 978-3-540-88681-5 978-3-540-88682-2. DOI: 10.1007/978-3-540-88682-2_24.

[11] L. Kantorovich. "On the transfer of masses (in russian)". In: *Doklady Akademii Nauk* 37.2 (1942), pp. 227–229.

[12] Matt J. Kusner et al. "From Word Embeddings to Document Distances". In: *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. Lille, France: JMLR.org, 2015, pp. 957–966.

[13] Gaspard Monge. *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale, 1781.

[14] A. Nemirovski et al. "Robust Stochastic Approximation Approach to Stochastic Programming". In: *SIAM Journal on Optimization* 19.4 (Jan. 2009), pp. 1574–1609. ISSN: 1052-6234, 1095-7189. DOI: 10.1137/070704277.

[15] O. Pele and M. Werman. "Fast and robust Earth Mover's Distances". In: *2009 IEEE 12th International Conference on Computer Vision*. 2009 IEEE 12th International Conference on Computer Vision. Sept. 2009, pp. 460–467. DOI: 10.1109/ICCV.2009.5459199.

[16] Gabriel Peyré and Marco Cuturi. "Computational Optimal Transport". In: *arXiv:1803.00567 [stat]* (Mar. 1, 2018). arXiv: 1803.00567.

[17] B. T. Polyak and A. B. Juditsky. "Acceleration of Stochastic Approximation by Averaging". In: *SIAM Journal on Control and Optimization* 30.4 (July 1992), pp. 838–855. ISSN: 0363-0129, 1095-7138. DOI: 10.1137/0330046.

[18] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. "The Earth Mover's Distance As a Metric for Image Retrieval". In: *Int. J. Comput. Vision* 40.2 (Nov. 2000), pp. 99–121. ISSN: 0920-5691. DOI: 10.1023/A:1026543900054.

[19] Mark Schmidt, Nicolas Le Roux, and Francis Bach. "Minimizing Finite Sums with the Stochastic Average Gradient". In: *arXiv:1309.2388 [cs, math, stat]* (Sept. 10, 2013). arXiv: 1309.2388.

[20] Vivien Seguy et al. "Large-Scale Optimal Transport and Mapping Estimation". In: *arXiv:1711.02283 [stat]* (Nov. 6, 2017). arXiv: 1711.02283.

[21] R. Sharathkumar and Pankaj K. Agarwal. "A near-linear time ε-approximation algorithm for geometric bipartite matching". In: *Proceedings of the 44th symposium on Theory of Computing - STOC '12*. the 44th symposium. New York, New York, USA: ACM Press, 2012, p. 385. ISBN: 978-1-4503-1245-5. DOI: 10.1145/2213977.2214014.

[22] Marcos Slomp et al. "GPU-based SoftAssign for maximizing image utilization in photomosaics". In: *International Journal of Networking and Computing* 1.2 (2011), pp. 211–229.

[23] Justin Solomon et al. "Convolutional wasserstein distances: efficient optimal transportation on geometric domains". In: *ACM Transactions on Graphics* 34.4 (July 27, 2015), 66:1–66. ISSN: 07300301. DOI: 10.1145/2766963.

[24] Günter Wyszecki and W. S. Stiles. *Color science: concepts and methods, quantitative data, and formulae*. Wiley classics library ed. Wiley classics library. New York: John Wiley & Sons, 2000. 950 pp. ISBN: 978-0-471-39918-6.