Report for CSE 5339 2018 — (OTMLSA)
Optimal Transport in Machine Learning and Shape Analysis

Stochastic Optimization for Large-scale Optimal Transport (OT)

Sixiong You

04/24/2018

1 Introduction

Optimal transport (OT) problem is the problem of finding the optimal map of transportation between two distributions, and OT defines a powerful framework to compare probability distributions in a geometrically faithful way. [1]. Due to its generality, OT has been applied in many fields of computer sciences, such as the classification of multi-label outputs [2], bag of word-embedding [3] and reflectance interpolation, color transfer, and geometry processing [4].

In Previous works, many different methods had been proposed to solve OT problem. In 1942, Kantorovitch invented Kantorovitch formulation to compute OT problem [7], which can be solved with Network flow solvers [8]. Recently, to smooth the distance estimation of OT problem and make the OT problem convex and more stable, some regularized methods were proposed, such as entropic regularization and group lasso.[6],[9]. Based on entropic regularization, Sinkhorn-Knopp algorithm was proposed to solve this problem [6]. However, these methods are purely discrete and cannot cope with continuous densities. In previous work, the only known class of methods that can overcome this limitation are so-called semi-discrete solvers [10]. In addition, the practical impact of OT is still limited because of its computational burden. Therefore, to deal with the large-scale OT problems, this report will introduce a stochastic optimization for large-scale optimal transport [5].

To be specific, three kinds of stochastic optimization methods are introduced to cope with three possible settings in this report:

1. For discrete OT, which compares a discrete measure with another discrete measure, the stochastic average gradient (SAG) method is applied.

2. For semi-discrete OT, which compares a discrete measure with a continuous measure, the stochastic gradient descent (SGD) method is applied.

3. For continuous OT, which compares a continuous measure with another continuous measure, via making use of an expansion of the dual variables in a reproducing kernel Hilbert space (RKHS) is applied.

The following of this report will be divided into 4 sections. In section 2, the formulation of OT and the formulation of stochastic optimization are introduced, In section 3 , the background of stochastic algorithm is introduced firstly, then the discrete optimal transport is introduced and results are presented, semi-discrete optimal transport and continuous optimal transport is introduced respectively.

2 Problem Formulation

2.1 The definition of joint probability measures and Kullback-Leible divergence

In the following we consider two metric spaces $\mathcal{X}$ and $\mathcal{Y}$. We denote by $\mathcal{M}_+^1(\mathcal{X})$ the set of positive Radon probability measures on $\mathcal{X}$. Let $\mu \in \mathcal{M}_+^1(\mathcal{X})$, $\nu \in \mathcal{M}_+^1(\mathcal{Y})$, then define the set of joint probability measures on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν as

$$\Pi(\mu,\nu) := \{\pi \in \mathcal{M}_+^1(\mathcal{X}); \forall (A,B) \subset \mathcal{X} \times \mathcal{Y}, \pi(A \times \mathcal{Y}) = \mu(A), \pi(\mathcal{X} \times B) = \nu(B)\}$$

To obtain the entropic regularization of OT problem, the Kullback-Leibler divergence between joint probabilities is defined as

$$\forall (\pi,\xi) \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})^2,$$

$$\mathbf{KL}(\pi|\xi) := \int_{\mathcal{X}\times\mathcal{Y}} (log(\frac{\mathrm{d}\pi}{\mathrm{d}\xi}(x,y)) - 1)\mathrm{d}\xi(x,y)$$

where $\frac{\mathrm{d}\pi}{\mathrm{d}\xi}$ is the relative density of π with respect to ξ. As the Kullback-Leibler divergence is also called the relative entropy of two different measures, that means when π does not have a density with respect to ξ, the $\mathbf{KL}(\pi|\xi) := +\infty$. In addtion, element-wise multiplication of vectors is denoted by $\odot$ and K^T denotes the transpose of a matrix K. Moreover, we also denote that $\mathbb{1}_N = (1,...,1)^T \in \mathbb{R}^N$ and $\circledcirc_N = (0,...,0)^T \in \mathbb{R}^N$. Besides, The Dirac measure at point x is δ_x. For a set C, $\iota_C(x) = 0$ if $x \in C$ and $\iota_C(x) = +\infty$ otherwise.

2.2 The problem formulation of optimal transport

We denote that $\forall (\mu,\nu) \in \mathcal{M}_+^1(\mathcal{X}) \times \mathcal{M}_+^1(\mathcal{Y})$, then the kantorovich formulation of OT can be written as

$$minimize \quad \gamma \mapsto \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y)$$

here $c \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})$ and $c(x,y)$ is the cost to transport a unit of mass from x to y. This c c is typically application-dependent, and reflects some prior knowledge on the data to process. Then by considering of Kullback-Leibler divergence, the OT can be written in a single convex optimization problem as

$$\mathbf{W}_\varepsilon(\mu,\nu) := \min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)\mathrm{d}\pi(x,y) + \varepsilon\mathbf{KL}(\pi|\mu\otimes\nu). \quad (\mathcal{P}_\varepsilon)$$

In addition, for any $c \in \mathcal{C}(\mathcal{X} \times \mathcal{Y})$, we define the following constraint set

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

and define the approximation of its indicator function as

$$\iota_{U_c}^{\varepsilon}(\mu,\nu) := \begin{cases} \iota_{U_c}(\mu,\nu) & \text{if} \quad \varepsilon = 0 \\ \varepsilon \int_{\mathcal{X}\times\mathcal{Y}} \exp(\frac{u(x)+v(y)-c(x,y)}{\varepsilon})\mathrm{d}\mu(x)\mathrm{d}\nu(y) & \text{if} \quad \varepsilon > 0 \end{cases}$$

For $\forall v \in \mathcal{C}(\mathcal{Y})$, $\forall x \in \mathcal{X}$we define the approximation of c-transform as

$$v^{c,\varepsilon}(x) := \begin{cases} \iota_{U_c}(\mu,\nu) & \text{if} \quad \varepsilon = 0 \\ -\varepsilon log(\int_{\mathcal{Y}} \exp(\frac{v(y)-c(x,y)}{\varepsilon})\mathrm{d}\nu(y)) & \text{if} \quad \varepsilon > 0 \end{cases}$$

Based on these definitions, to realize the application of stochastic optimization methods, two dual problems are described. First, based on Fenchel-Rockafellar's dual theorem, the dual formulation of original OT problem $\mathcal{P}_\varepsilon$ can be expressed as [5]

$$\mathbf{W}_\varepsilon(\mu,\nu) = \max_{v\in\mathcal{C}(\mathcal{Y}),u\in\mathcal{C}(\mathcal{X})} F_\varepsilon(u,v) := \int_{\mathcal{X}} u(x)\mathrm{d}\mu(x) + \int_{\mathcal{Y}} v(x)\mathrm{d}\nu(y) - \iota^{\varepsilon}_{U_c}(\mu,\nu). \quad (\mathcal{D}_\varepsilon)$$

Then by solving $\frac{\partial F_\varepsilon(u,v)}{\partial u} = 0$, we can obtain that for $\varepsilon > 0$

$$\frac{\partial F_\varepsilon(u,v)}{\partial u} = u(x) - \int_{\mathcal{Y}} \exp(\frac{v(y)-c(x,y)}{\varepsilon})\mathrm{d}\nu(y)\exp(\frac{u(x)}{\varepsilon})u(x) = 0 \Rightarrow$$

$$u(x) = -\varepsilon log(\int_{\mathcal{Y}} \exp(\frac{v(y)-c(x,y)}{\varepsilon})\mathrm{d}\nu(y)) = v^{c,\varepsilon}(x)$$

for $\varepsilon = 0$, as for any $c \in \mathcal{C}(\mathcal{X}\times\mathcal{Y})$, we have $u(x)+v(y) \leqslant c(x,y)$, so there is $u(x) \leqslant c(x,y)-v(y)$, from which we can obtain the approximation of $u(x) \approx \min_{y\in\mathcal{Y}} c(x,y) - v(y)$, therefore, for $\varepsilon \geqslant 0$, we can obtain $u(x) = v^{c,\varepsilon}(x)$, and plugging this expression back in $(\mathcal{D}_\varepsilon)$, we can obtain the semi-dual formulation of OT

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

To apply stochastic algorithm to dual and semi-dual problems, the equations $(\mathcal{D}_\varepsilon)$ and $(\mathcal{S}_\varepsilon)$ must be expressed with expectations, which can be written as

$$\forall\varepsilon > 0, F_\varepsilon(u,v) = \mathbb{E}_{X,Y}[f_\varepsilon(X,Y,u,v)] \quad and \quad \forall\varepsilon \geqslant 0, H_\varepsilon(u,v) = \mathbb{E}_X[h_\varepsilon(X,v)]$$

in which X and Y are independent and distributed according to μ and ν. Therefore,when $\varepsilon > 0$,

$$\begin{aligned}
F_\varepsilon(u,v) &:= \int_{\mathcal{X}} u(x)\mathrm{d}\mu(x) + \int_{\mathcal{Y}} v(x)\mathrm{d}\nu(y) - \iota^{\varepsilon}_{U_c}(\mu,\nu) \\
&= \int_{\mathcal{X}} 1\times u(x)\mathrm{d}\mu(x) + \int_{\mathcal{Y}} 1\times v(x)\mathrm{d}\nu(y) - \iota^{\varepsilon}_{U_c}(\mu,\nu) \\
&= \int_{\mathcal{X}}\int_{\mathcal{Y}} \mathrm{d}\nu(y)\times u(x)\mathrm{d}\mu(x) + \int_{\mathcal{Y}}\int_{\mathcal{X}} \mathrm{d}\mu(x)\times v(y)\mathrm{d}\nu(y) - \iota^{\varepsilon}_{U_c}(\mu,\nu) \\
&= \int_{\mathcal{X}\times\mathcal{Y}} u(x)\mathrm{d}\nu(y)\mathrm{d}\mu(x) + \int_{\mathcal{X}\times\mathcal{Y}} v(y)\mathrm{d}\nu(y)\mathrm{d}\mu(x) - \varepsilon\int_{\mathcal{X}\times\mathcal{Y}} \exp(\frac{u(x)+v(y)-c(x,y)}{\varepsilon})\mathrm{d}\mu(x)\mathrm{d}\nu(y)
\end{aligned}$$

Therefore, we can define

$$\forall\varepsilon > 0, f_\varepsilon(X,Y,u,v)] := u(x) + v(y) - \varepsilon exp(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}),$$

Similarly, we can define that

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

Then based on above expectation maximization problem, three different stochastic algorithms are introduced for three kinds of situations.

3 Discrete Optimal Transport

3.1 The introduction of stochastic algorithm

In this report, the adopted stochatic algorithms are based on stochastic gradient descent (SGD), so this algorithm is simply introduced. The fundamental of SGD is that using the gradient of one stochastic term to replace the full gradient to maximize the objective in each iteration. For example, for an optimized problem

$$J = \min_{\omega} Q(\omega) = \min_{\omega} \frac{1}{n} \sum_{i=1}^{n} Q_i(\omega)$$

in which J is the objective function and ω is the independent variable, a standard gradient descent method in each iteration woruld be expressed as

$$\omega := \omega - \eta \nabla Q(\omega) = \omega - \eta \frac{1}{n} \sum_{i=1}^{n} \nabla Q_i(\omega)$$

where η is a step size, ∇Q_i is the gradient of Q_i. This method is also called batch gradient descent method. However, expensive evaluations of all gradients is needed to evaluate the sum-gradient. To reduce the computational cost at every iteration, SGD samples a subset of all gradients at every step, which is very effective for large-scale machine learning problems. To further improve the performance of SGD, the stochastic average gradient (SAG) is proposed, the convergence rate is improved from $O(1/\sqrt{k})$ to $O(1/k)$ by keeping and averaging the previous stochastic gradient values [11].

3.2 SGD for discrete optimal transport

Assuming μ and ν are discrete measures, and the cost matrix $c \in \mathbb{R}^{I \times J}$ defined by $c_{i,j} = c(x_i, y_j)$, then the problems $(\mathcal{P}_\varepsilon)$,$(\mathcal{D}_\varepsilon)$ and $(\mathcal{S}_\varepsilon)$ can be reformulated as

$$\begin{aligned} W_\varepsilon(u,v) &= \min_{\pi \in \mathbb{R}^{I \times J}} \{\sum_{i,j} \mathbf{c}_{i,j}\pi_{i,j} + \varepsilon \sum_{i,j} (log\frac{\pi_{i,j}}{\mu_i \nu_j} - 1)\pi_{i,j}; \pi \mathbb{1}_J = \mu, \pi^T \mathbb{1}_I = \nu\} \quad (\mathcal{P}_{\varepsilon a}) \\ &= \max_{u \in \mathbb{R}^I, v \in \mathbb{R}^J} \sum_i \mathbf{u}_i \mu_i + \sum_j \mathbf{v}_j \nu_j - \varepsilon \sum_{i,j} exp(\frac{u_i + v_j - c_{i,j}}{\varepsilon})\mu_i \nu_j, (for \ \ \varepsilon > 0) \quad (\mathcal{D}_{\varepsilon a}) \\ &= \max_{v \in \mathbb{R}^J} H_{\varepsilon a} = \sum_{i \in I} h_{\varepsilon a}(x_i, \mathbf{v})\mu_i \quad (\mathcal{S}_{\varepsilon a}) \end{aligned}$$

$$h_{\varepsilon a}(x, \mathbf{v}) = \sum_{j \in J} \mathbf{v}_j \nu_j + \begin{cases} \min_j (c(x, y_j) - \mathbf{v}_j) & \text{if} \quad \varepsilon = 0 \\ -\varepsilon log(\sum_{j \in J} \exp(\frac{v_j - c(x, y_j)}{\varepsilon})\nu_j) - \varepsilon & \text{if} \quad \varepsilon > 0 \end{cases}$$

To solve this optimization problem, the SAG is proposed. For SAG, by setting the step size of SAG, and simple initialization (setting the initial guess of gradients and output to be zero), this SAG can start to work. Via sampling finite gradient $\mathbf{g}_i$ in iteration i, SAG can update the gradient $\mathbf{d}$ based on previous gradient data $\mathbf{d} \leftarrow \mathbf{d} + \mathbf{g_i}$. Finally, SAG can approach the optimal solution of OT by calculating $\mathbf{v} \leftarrow \mathbf{v} + C\mathbf{d}$ in each iteration. When the termination condition is reached, the algorithm is completed. To verify the effectiveness of the proposed SAG, the bags of word-embeddings is adopted as numerical illustrations and Sinkhorn is used for comparison. Results show that the SAG can be twice faster than Sinkhorn on average while the same parallel properties can be obtained for both methods [5].

4 Semi-Discrete Optimal Transport

For semi-discrete optimal transport, which compares a discrete measure with a continuous measure. Therefore, assuming μ is continuous measure, which means the semi-dual problem $(\mathcal{S}_{\varepsilon a})$ is still applicable to this case. Because in $(\mathcal{S}_{\varepsilon a})$, only μ is not needed to be discrete. However, as for SAG, the update of sampling gradient $\mathbf{g}_i$ needs μ_i to be discrete, so SAG cannot be used for semi-discrete optimal transport. To solve this problem, the average SGD is adopted which has no requirements on μ[12]. For average SGD, in each step, the average SGD will obtain the sampling gradients by calculating $\frac{C}{\sqrt{k}}\nabla_v h_{\varepsilon a}(x_k, \mathbf{v}')$, and by averaging all existing gradients via $\mathbf{v} \leftarrow \frac{1}{k}\mathbf{v}' + \frac{k-1}{k}\mathbf{v}$, this algorithm can approach the optimal solution of OT with the convergence rate of $O(1/\sqrt{k})$. Finally a numerical illustration is adopted for testing the effectiveness of average SGD. By comparing the resutls of SAG and SGD for semi-discrete illustrations, we can conclude that due to the estimation error from discretization of μ, SAG cannot converges to the correct solution of semi-discrete optimal transport while SGD can converge to the correct solution of semi-discrete optimal transport[5].

5 Continuous Optimal Transport

For two RKHS $\mathcal{H}$ and $\mathcal{G}$ defined on $\mathcal{X}$ and $\mathcal{Y}$, with kernels κ and ζ, associated with norms $\|.\|_{\mathcal{H}}$ and norms $\|.\|_{\mathcal{G}}$. Based on the expectations of $(\mathcal{D}_\varepsilon)$, via applying SGD to this problem, start from $u_0 = 0$ and $v_0 = 0$, we can have

$$(u_k, v_k) := (u_{k-1}, v_{k-1}) + \frac{C}{\sqrt{k}}\nabla f_\varepsilon(x_k, y_k, u_{k-1}, v_{k-1}) \in \mathcal{H} \times \mathcal{G}$$

Proposition 5.1. The (u_k, v_k) defined above satisfy

$$(u_k, v_k) := \sum_{i=1}^{k} \alpha_i(\kappa(., x_i), \zeta(., y_i)), \text{ in which } \alpha_i := \Pi_{B_r}(\frac{C}{\sqrt{i}}(1-e(\frac{u_{i-1}(x_i) + v_{i-1}(y_i) - c(x_i, y_i)}{\varepsilon})))$$

where $(x_i, y_i)_{i=1,..,k}$ are samples from $\mu \otimes \nu$ and Π_{B_r} is the projection on the centered ball of radius r.

Then, combing with SGD, the kernel SGD for continuous OT can be obtained. In each step, through sampling x_k from μ and y_k from ν, we can update u_{k-1}, v_{k-1} and α_k in step k through the following equation: $u_{k-1}(x_k) := \sum_{i=1}^{k-1} \alpha_i \kappa(x_k, x_i)$,$v_{k-1}(y_k) := \sum_{i=1}^{k-1} \alpha_i \zeta(y_k, y_i)$,$\alpha_k := \frac{C}{\sqrt{k}}(1-e(\frac{u_{k-1}(x_k)+v_{k-1}(y_k)-c(x_k,y_k)}{\varepsilon}))$. When the terminal condition is reached, the final (α_k, x_k, y_k) can be obtained. From numerical results, we can find that the convergence of kernel SGD is slow,and when μ has more mass, the kernel SGD can converge faster, which means the value of u has the greatest impact in F_ε[5].

References

[1] MONGE, G. "Memoire sur la theorie des deblais et des remblais." Histoire de l'Académie Royale des Sciences de Paris. 1781. https://ci.nii.ac.jp/naid/10018386702/en/.

[2] Frogner, Charlie and Zhang, Chiyuan and Mobahi, Hossein and Araya, Mauricio and Poggio, Tomaso A. "Learning with a Wasserstein Loss." In Advances in Neural Information Processing Systems, pp. 2053-2061. 2015.http://papers.nips.cc/paper/5679-learning-with-a-wasserstein-loss.pdf.

[3] Kusner, Matt and Sun, Yu and Kolkin, Nicholas and Weinberger, Kilian. "From word embeddings to document distances." In International Conference on Machine Learning, pp. 957-966. 2015.http://proceedings.mlr.press/v37/kusnerb15.pdf.

[4] Solomon, Justin and de Goes, Fernando and Peyré, Gabriel and Cuturi, Marco and Butscher, Adrian and Nguyen, Andy and Du, Tao and Guibas, Leonidas. "Convolutional Wasserstein Distances: Efficient Optimal Transportation on Geometric Domains." ACM Transactions on Graphics (TOG) 34, no. 4 (2015): 66. http://doi.acm.org/10.1145/2766963.

[5] Genevay, Aude and Cuturi, Marco and Peyré, Gabriel and Bach, Francis. "Stochastic Optimization for Large-scale Optimal Transport." In Advances in Neural Information Processing Systems, pp. 3440-3448. 2016. http://papers.nips.cc/paper/6566-stochastic-optimization-for-large-scale-optimal-transport.pdf.

[6] Cuturi, Marco. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport." In Advances in neural information processing systems, pp. 2292-2300. 2013. http://papers.nips.cc/paper/4927-sinkhorn-distances-lightspeed-computation-of-optimal-transport.pdf.

[7] Kantorovitch, L. V. "On the translocation of masses." Management Science 5, no. 1 (1958): 1-4. https://ci.nii.ac.jp/naid/10025352868/en/.

[8] R. Burkard, M. Dell'Amico, and S. Martello."Assignment Problems." revised reprint. Vol. 125. Siam, 2009. http://dl.merc.ac.ir/handle/Hannan/8512.

[9] Courty, Nicolas and Flamary, Rémi and Tuia, Devis and Rakotomamonjy, Alain. "Optimal Transport for Domain Adaptation." IEEE transactions on pattern analysis and machine intelligence 39, no. 9 (2017): 1853-1865. https://hal.archives-ouvertes.fr/hal-01170705.

[10] Aurenhammer, F.and Hoffmann, F.and Aronov, B. "Minkowski-Type Theorems and Least-Squares Clustering." Algorithmica 20, no. 1 (1998): 61-76. https://doi.org/10.1007/PL00009187.

[11] Schmidt, Mark. "Minimizing finite sums with the stochastic average gradient." Mathematical Programming 162, no. 1-2 (2017): 83-112 https://open.library.ubc.ca/cIRcle/collections/48630/items/1.0044624.

[12] Polyak, B. T. "Acceleration of stochastic approximation by averaging." SIAM Journal on Control and Optimization 30, no. 4 (1992): 838-855. https://ci.nii.ac.jp/naid/80006605715/en/.