Onlicorne: optimal transportation distances from sample streams

Anonymous Authors[1]

Abstract

Optimal Transport (OT) distances are now routinely used as loss functions in ML tasks. Yet, computing OT distances between arbitrary (i.e. not necessarily discrete) probability distributions remains an open problem. This paper introduces a new online estimator of entropy-regularized OT distances between two such arbitrary distributions. It uses streams of samples from both distributions to iteratively enrich a non-parametric representation of the transportation plan. Compared to the classic Sinkhorn algorithm, our method leverages new samples at each iteration, which enables a consistent estimation of the true regularized OT distance. We cast our algorithm as a block-convex mirror descent in the space of positive distributions, which enables a theoretical analysis of its convergence. We numerically illustrate the performance of our method in comparison with concurrent approaches.

1. Introduction

Optimal transport (OT) distances are fundamental in statistical learning, both as a tool for analyzing the convergence of various algorithms (Canas & Rosasco, 2012; Dalalyan & Karagulyan, 2019), and as a data-dependent term for tasks as diverse as supervised learning (Frogner et al., 2015), unsupervised generative modeling (Martin Arjovsky, 2017) or domain adaptation (Courty et al., 2016). OT lifts a given distance over data points living in space $\mathcal{X}$ into a distance on the space $\mathcal{P}(\mathcal{X})$ of probability distributions over this data space $\mathcal{X}$, and we refer to the monograph (Santambrogio, 2015) for a detailed mathematical treatement. This distance has many favorable geometrical properties, in particular it allows one to compare distributions having disjoint supports. Computing OT distance is usually performed by sampling the input distributions and solving a discretized linear program, the so-called Kantorovich formuation (Kantorovich, 1942). This approach is numerically costly and is difficult to apply in typical ML scenario where data points are streamed in an online manner. The goal of this paper is to develop an efficient online method which addresses these issues by adapting Sikhorn's algorithm to an online setting.

Regularized OT. To alleviate both these computational and statistical burdens, it is common to regularize the associated linear program. The most successful approach in this direction is to use an entropic barrier penalty. When dealing with discrete distributions, this regularization can be solved numerically using Sinkhorn-Knopp's matrix balancing algorithm (Sinkhorn, 1964; Sinkhorn & Knopp, 1967). This approach was pushed forward for ML applications by Cuturi (Cuturi, 2013) who emphasized both its parallelism and its smoothing effects, which makes this approach a perfect fit when training ML model through back-propagation. This method estimates the OT distance in two distinct phases: one draws samples and evaluate a pairwise distance matrix in the first phase ; one balances this distance matrix using Sinkhorn-Knopp iterations in the second phase, thereby obtaining a discretized transportation plan and distance.

This approach offers many advantage over the direct solution of a linear solver. First it computes an ε-accurate approximation of OT in $O(n^2/\varepsilon^3)$ for a number n of samples (Altschuler et al., 2017) (in contrast to the $O(n^3)$ complexity for an exact solution). Second, the optimal value of the regularized problem does not suffers from the curse of dimensionality (Genevay et al., 2019), since the average error using n random samples decay likes $O(\varepsilon^{-d/2}/\sqrt{n})$, in sharp contrast with the slow $O(1/n^{1/d})$ error decay of OT (Weed et al., 2019). This regularized value can be debiased to define the so-called Sinkhorn divergence (Feydy et al., 2019).

This two step (sample and then compute) approach is however not able to introduce exact OT loss in learning problem, which often rather operate by accessing new samples during the iterations. A cheap fix is to use Sinkhorn over mini-batches (see for instance (Genevay et al., 2018) for an application to GANs), but this introduce a strong bias which can be unacceptable, especially in high dimension (see (Fatras et al., 2019) for a mathematical analysis).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Extending OT computations to arbitrary distributions (possibly having continuous densities) without relying on such a fixed a priori sampling is an emerging topic of interest. A special case is the semi-discrete setting, where one of the two distributions is discrete. Without regularization, over an Euclidean space, this can be solved efficiently using the computation of Voronoi-like diagrams (Mérigot, 2011). This idea can be extended to entropic-regularized OT (Cuturi & Peyré, 2018), and can also be coupled with stochastic optimization method (Genevay et al., 2016) to tackle high dimensional problems (see also (Staib et al., 2017) for an extension to the computation of Wasserstein barycenters).

When dealing with arbitrary continuous densities, which are accessed through a stream of random sample, the challenge is to approximate the (continuous) dual variables using parametric functions. For application to generative model fitting, one can use deep networks, which leads to alternative formulation to Generative Adversarial Networks (GANs) (Martin Arjovsky, 2017) (see also (Seguy et al., 2018) for an extension to the estimation of transportation maps). There is however no theoretical guarantees for this type of dual approximations, due to the non-convexity of the resulting optimization problem. The only mathematically rigorous approach in this direction is when using expansions in a reproducing Hilbert space (Genevay et al., 2016). Our paper proposes a different takes on this question, using an alternative type of expansions, which corresponds to an extension of the discrete Sinkhorn algorithm to the continuous online setting.

Contribution. In this paper, we show how mingling together the sampling and the optimization phases can be beneficial to quickly estimate OT distances for ML problems. Our approach relies on three observations. First, Sinkhorn iterations can be rewritten as a block convex mirror descent on the space of positive distributions. This formulation is valid in the discrete and continuous setting. Second, we can modify these iterations to rely on realizations $\hat{\alpha}_t$, $\hat{\beta}_t$ of the input distributions α and β, renewed at each iteration t. Finally, we can represent the iterates produced by such approximations in a space of mixtures of simple functions. Those iterates are a simple transformation of the potentials in the Sinkhorn optimization problem. This corresponds to the following contributions:

- We introduce a new *online Sinkhorn* algorithm. It produces a sequence of estimates $(\hat{w}_t)_t \in \mathbb{R}$ and of transportation plans $\hat{\pi}_t \in \mathcal{P}(\mathcal{X} \times \mathcal{X})$ TODO: need to introduce this, using two stream of renewed samples $\hat{\alpha}_t = \sum_{i=1}^n \delta_{x_i^t}$, $\hat{\beta}_t = \sum_{i=1}^n \delta_{y_i^t}$, where x_i^t and y_i^t are sampled from α and β.
- We show that those estimations are consistent, in the sense that $\hat{w}_t \to \mathcal{W}_{C,\varepsilon}(\alpha, \beta)$, and $\hat{\pi}_t \to \pi^\star$ TODO: weak convergence?.
- We empirically demonstrate that our algorithm permits a faster estimation of optimal transportation distances for discrete distributions, and a convincing estimation of OT distances between *continous* distributions.

2. Background: optimal transport distances

We recall the definition of optimal transport distances between arbitrary distributions (i.e. not necessarily discrete), then review how these are estimated using finite samples.

2.1. Optimal transport distances and algorithms

Wasserstein distances. We consider a complete metric space $(\mathcal{X}, d)$ (assumed to be compact for simplicity), equipped with a continuous cost function $C(x, y) \in \mathbb{R}$ for any $(x, y) \in \mathcal{X}^2$ (assumed to be symmetric also for simplicity). Optimal transport lifts this *ground cost* into a cost between probability distributions over the space $\mathcal{X}$.

The Wasserstein cost between two probability distribution $(\alpha, \beta) \in \mathcal{P}(\mathcal{X})$ is defined as the minimal cost required to move each element of mass of α to each element of mass of β. It rewrites as the solution of a linear problem (LP) over the set of transportation plans (which are probability distribution π over $\mathcal{X} \times \mathcal{X}$)

$$\mathcal{W}_C(\alpha, \beta) \triangleq \min_{\pi \in \mathcal{P}(\mathcal{X}^2)} \left\{ \langle C, \pi \rangle \ : \ \pi_1 = \alpha, \pi_2 = \beta \right\}, \quad (1)$$

where we denote $\langle C, \pi \rangle \triangleq \int C(x, y) \mathrm{d}\pi(x, y)$. Here, $\pi_1 = \int_{y \in \mathcal{X}} \mathrm{d}\pi(\cdot, y)$ and $\pi_2 = \int_{x \in \mathcal{X}} \mathrm{d}\pi(x, \cdot)$ are the first and second marginals of the transportation plan π. When $C = d^p(x, y)$ is the p^{th} power of the ground distance, with $p \geqslant 1$, then $\mathcal{W}_C^{\frac{1}{p}}$ is itself a distance over $\mathcal{P}(\mathcal{X})$, whose associated topology is the one of the convergence in law (Santambrogio, 2015).

Entropic regularization and Sinkhorn algorithm. The solutions of (2.1) can be approximated by a strictly convex optimisation problem, where an entropic term is added to the linear objective to force curvature. The so-called Sinkhorn cost is then

$$\mathcal{W}_{C,\varepsilon}(\alpha, \beta) \triangleq \min_{\substack{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) \\ \pi_1 = \alpha, \pi_2 = \beta}} \langle C, \pi \rangle + \varepsilon \mathrm{KL}(\pi | \alpha \otimes \beta), \quad (2)$$

where the Kulback-Leibler divergence is defined as $\mathrm{KL}(\pi | \alpha \otimes \beta) \triangleq \int \log(\frac{\mathrm{d}\pi}{\mathrm{d}\alpha \mathrm{d}\beta}) \mathrm{d}\pi$ (which is thus equal to the mutual information of π). This can be shown to approximate up to an $\varepsilon \log(\varepsilon)$ error $\mathcal{W}_C(\alpha, \beta)$ (see (Genevay et al., 2019)), which is recovered in the limit $\varepsilon = 0$. The regularized problem (2.1) admits a dual form, which is a maximization problem over the space of continuous *potential* function:

$$\max_{f,g \in \mathcal{C}(\mathcal{X})} \langle f,\, \alpha\rangle + \langle g,\, \beta\rangle - \varepsilon\langle \alpha\otimes\beta,\, \exp(\frac{f\oplus g - C}{\varepsilon})\rangle + \varepsilon, \tag{3}$$

where $\langle f,\, \alpha\rangle \triangleq \int f(x)\mathrm{d}\alpha(x)$ and $(f\oplus g - C)(x) \triangleq f(x) + g(y) - C(x,y)$. We refer to (?) for more details on this problem.

The major interest in this regularization is because it can be efficiently solved by alternated maximization on (2.1). This method, at iteration t

$$f_{t+1}(\cdot) = -T_{C,\varepsilon}(g_t, \beta),\; g_{t+1}(\cdot) = -T_{C,\varepsilon}(f_{t+1}, \alpha), \tag{4}$$

$$\text{where} \quad T_C(h,\mu) \triangleq \int_{y\in\mathcal{X}} \exp(\frac{h(y) - C(\cdot,y)}{\varepsilon})\mathrm{d}\mu(y),$$

TODO: j'ai suppose $C = C^\top$ symetrique pour simplifier The operation $h \mapsto T_C(h,\mu)$ maps a continuous function to another continuous function, and is a smooth approximation of the celebrated C-transform of OT (Santambrogio, 2015), we thus refers to it as a *soft C-transform*. The notation $f_t(\cdot)$ emphasizes the fact that f_t and g_t are continuous functions.

It can be shown that $(f_t)_t$ and $(g_t)_t$ converge in $(\mathcal{C}(\mathcal{X}), \|\cdot\|_{\text{var}})$ to a solution $(f^\star, g^\star)$ of (2.1), where $\|f\|_{\text{var}} = \max_x f(x) - \min_x f(x)$ is the so-called variation norm. Convergence is due to the strict contraction of the operators $T_C(\cdot, \beta)$ and $T_C(\cdot,\alpha)$ in the space $(\mathcal{C}(\mathcal{X}), \|\cdot\|_{\text{var}})$ (?).

2.2. Estimating OT distances with realizations

Iterations (2.1) cannot be implemented when dealing with generic distributions (α,β), because it involves continuous functions (f_t, g_t). When the input distribution are discrete (or equivalently that $\mathcal{X}$ is a finite set) then these function can be stored using discrete sample, and algorithm (2.1) is equivalent to the celebrated Sinkohrn's algorithm (Sinkhorn, 1964; Sinkhorn & Knopp, 1967), which is often implemented over the scaling variable $(e^{f_t/\varepsilon}, e^{g_t/\varepsilon})$. More precisely, when dealing with empirical distibutions $\hat\alpha = \sum_{i=1}^n a_i\delta_{x_i}$ and $\hat\beta = \sum_{i=1}^n a_i\delta_{x_i}$, then the approximation of $\mathcal{W}_{C,\varepsilon}(\alpha,\beta)$ using Sinkohrn iterations (2.1) compute $u_t \triangleq (e^{f_t(x_i)/\varepsilon})_{i=1}^n, v_t \triangleq (e^{g_t(y_i)/\varepsilon})_{i=1}^n$ as

$$u_{t+1} = \frac{1}{K(v_t \odot a)} \quad \text{and} \quad v_{t+1} = \frac{1}{K^\top(u_{t+1}\odot b)}$$

where $K = (e^{-\frac{C(x_i,y_j)}{\varepsilon}})_{i,j=1}^n \in \mathbb{R}^{n\times n}$. The algorithm thus operates in two phases: a first one, during which the kernel matrix K is computed (with a cost in $O(n^2)$), and a second one, during which it is balanced (each iteration having a cost in $O(n^2)$).

The goal of this paper is to go beyond this discrete setting, and handle generic distributions (possibly having continuous densities). In particular, our numerical scheme manipulates continuous functions though an adapted parameterization which is automatically refined during the iterations.

Sample complexity results. TODO: I already say this quickly in the intro, check if we remove this section Fortunately, the OT and Sinkhorn distances between two arbitrary distributions α and β can be approximated by the distance between discrete realizations $\hat\alpha_n = \frac{1}{n}\sum_i \delta_{x_i}$, $\hat\beta_n = \frac{1}{n}\sum_i \delta_{y_i}$, where $(x_i)_i$ and $(y_i)_i$ are i.i.d samples from α and β. Consistency holds, as $\mathcal{W}_{C,(\varepsilon)}(\hat\alpha_n, \hat\beta_n) \to \mathcal{W}_{C,(\varepsilon)}(\alpha,\beta)$, with a convergence rate in $\mathcal{O}(n^{-1/2})$ for Sinkhorn distances and $\mathcal{O}(n^{-1/d})$ for Wasserstein distances.

Bias in distance estimation. Although consistency is a reassuring result, the sample complexity of transport in high dimensions with low regularization remains high. For computational reasons, we cannot choose n to be much more than 10^5, which is not sufficient to ensure that $\mathcal{W}_{C,\varepsilon}(\alpha,\beta)$ is ε-close to $\mathcal{W}_C$ in the typical case where $d =?$ and $\varepsilon =?$.

We may wonder wether we can improve the estimation of $\mathcal{W}_C(\alpha,\beta)$ using several sets of samples $(\hat\alpha_n^t)_t$ and $(\hat\beta_n^t)$. Those should be of reasonable size to allow Sinkhorn estimation, and may for example come from a temporal stream. ()() have proposed to use the Monte-Carlo estimate $\hat{\mathcal{W}}(\alpha,\beta) = \frac{1}{T}\sum_{t=1}^T \mathcal{W}(\hat\alpha_n^t, \beta\alpha_n^t)$. However, this yields a wrong estimation as the distance $\mathcal{W}(\hat\alpha_n, \hat\beta_n)$ between discrete realizations is a *biased* estimator of $\mathcal{W}(\alpha,\beta)$:

$$\mathcal{W}(\alpha,\beta) \neq \mathbb{E}_{\hat\alpha_n\sim\alpha, \hat\beta_n\sim\beta}[\mathcal{W}(\hat\alpha_n, \hat\beta_n)].$$

Bias in gradients. In several applications, the distance $\mathcal{W}(\alpha,\beta)$ is used as a loss function. This is the case in generative modeling, when we parametrize α as the push-forward of some noise distribution μ through a neural network g_θ. We are then interested in computing the displacement gradient $\delta_\alpha\mathcal{W}(\alpha,\beta) \in \mathcal{P}(\mathcal{X})$, in order to train θ by backpropagation. This gradient turns out to be the spatial derivative $\nabla_x f^\star$ of the solution of (2.1). Yet, similarly, estimating this gradient through sampling is biased, as $f^\star(\alpha,\beta) \neq \mathbb{E}_{\hat\alpha_n\sim\alpha,\hat\beta_n\sim\beta}[f^\star(\hat\alpha_n,\hat\beta_n)]$.

3. OT distances from sample streams

We introduce a novel understanding of the Sinkhorn algorithm in this section, whence we derive average and online adaptations. We wish to construct an estimator of $\mathcal{W}(\alpha,\beta)$ from multiple sets of samples $(\hat\alpha_n^t)_t$ and $(\hat\beta_n^t)_t$. This estimator should successively use these samples to enrich a representation of the solution of (2.1), that may be arbitrary complex. $(\hat\alpha_n^t)_t$ and $(\hat\beta_n^t)_t$ may be seen as mini-batches within a training procedure, or as a temporal stream.

3.1. Unbiased Sinkhorn iterations

3.2. Offline mini-batch averaging

3.3. Online estimation of Sinkhorn distances

3.4. Bias-variance trade-offs

4. Analysis

4.1. Offline Sinkhorn averaging

Estimator $\langle \alpha, -\log \mathbb{E}[\exp(-\hat{f})]\rangle + \langle \beta, -\log \mathbb{E}[\exp(-\hat{g})]\rangle$

- estimator properties

4.2. Online Sinkhorn convergence

- slowed down Sinkhorn convergence
- Random iterated functions
- Combining both

4.3. Non-convex mirror descent

5. Experiments

5.1. Offline distance averaging

5.2. Online distance computations

5.3. Training generative models

References

Altschuler, J., Niles-Weed, J., and Rigollet, P. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pp. 1964–1974, 2017.

Canas, G. and Rosasco, L. Learning probability measures with respect to optimal transport metrics. In *Advances in Neural Information Processing Systems*, pp. 2492–2500, 2012.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9): 1853–1865, 2016.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pp. 2292–2300, 2013.

Cuturi, M. and Peyré, G. Semidual regularized optimal transport. *SIAM Review*, 60(4):941–965, 2018. doi: 10.1137/18M1208654. URL https://arxiv.org/abs/1811.05527.

Dalalyan, A. S. and Karagulyan, A. User-friendly guarantees for the langevin monte carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12): 5278–5311, 2019.

Fatras, K., Zine, Y., Flamary, R., Gribonval, R., and Courty, N. Learning with minibatch wasserstein: asymptotic and gradient properties. *arXiv preprint arXiv:1910.04091*, 2019.

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S., Trouvé, A., and Peyré, G. Interpolating between optimal transport and mmd using sinkhorn divergences. In *Proc. AISTATS'19*, 2019. URL https://arxiv.org/abs/1810.08278.

Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. Learning with a wasserstein loss. In *Advances in Neural Information Processing Systems*, pp. 2053–2061, 2015.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. In Lee, D. D., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Proc. NIPS'16*, pp. 3432–3440. Curran Associates, Inc., 2016. URL http://hal.archives-ouvertes.fr/hal-01321664.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *Proc. AISTATS'18*, pp. 1608–1617, 2018. URL https://arxiv.org/abs/1706.00292.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. Sample complexity of sinkhorn divergences. In *Proc. AISTATS'19*, 2019. URL https://arxiv.org/abs/1810.02733.

Kantorovich, L. On the transfer of masses (in russian). *Doklady Akademii Nauk*, 37(2):227–229, 1942.

Martin Arjovsky, Soumith Chintala, L. B. Wasserstein generative adversarial network. *Proc ICML'17*, 60(4):941–965, 2017.

Mérigot, Q. A multiscale approach to optimal transport. In *Computer Graphics Forum*, volume 30, pp. 1583–1592. Wiley Online Library, 2011.

Santambrogio, F. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Seguy, V., Damodaran, B. B., Flamary, R., Courty, N., Rolet, A., and Blondel, M. Large-scale optimal transport and mapping estimation. *Proc. ICLR*, 2018.

Sinkhorn, R. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.*, 35:876–879, 1964.

Sinkhorn, R. and Knopp, P. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Staib, M., Claici, S., Solomon, J. M., and Jegelka, S. Parallel streaming wasserstein barycenters. In *Advances in Neural Information Processing Systems*, pp. 2647–2658, 2017.

Weed, J., Bach, F., et al. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.