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, and as a data-dependent term for estimating data densities, e.g. using generative models. 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 $\mathcal{X}$. 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. This approach is numerically costly and is difficult to apply in typical ML scenario where data points are streamded 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.

OT in Machine Learning. Typical applications.

Computation bottleneck.

Statistical bottleneck.

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 (). This approach was pushed forward for ML applications by Cuturi () who emphasized both its parallelism and its smoothing effects, which makes this approach a perfect fit when training ML model through back-propagation. CITE 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.

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 (). This idea can be extended to entropic-regularized OT (), which can be coupled with stochastic optimization method.

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) (). TODO: CITER d'autre papiers similaire d'approximation continues? 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 (). Our paper pro-

[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.

poses a different takes on this question, using a 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 these two phases can be beneficial to quickly estimate OT distances. 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 two 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.

These observations allows us to propose the following material.

- 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 space $\mathcal{X}$, equiped with a distance function $C : \mathcal{X} \to \mathcal{X} \to \mathbb{R}$. Optimal transport lifts this *ground metric* distance into a distance between probability distributions over the space $\mathcal{X}$. The Wasserstein distance between α and β 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

$$\mathcal{W}_C(\alpha,\beta) = \min_{\substack{\pi\in\mathbb{P}(\mathcal{X}\times\mathcal{X})\\ \pi_1=\alpha,\pi_2=\beta}} \langle C,\, \pi\rangle,$$

where $\pi_1 = \int_{y\in\mathcal{X}} d\pi(\cdot, y)$ and $\pi_2 = \int_{x\in\mathcal{X}} d\pi(x,\cdot)$ are the first and second marginals of the transportation plan π. It can shown (ref) that $\mathcal{W}_C$ is a distance over $\mathcal{P}(\mathcal{X})$, that measures the convergence in law.

Entropic regularization and Sinkhorn algorithm. The solution of (**??**) can be approximated by a simpler optimisation problem, where an entropic term is added to the linear objective to force curvature. The so-called Sinkhorn distance

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

is indeed an ε-approximation of $\mathcal{W}_C(\alpha,\beta)$. The later problem admits a dual form, which is a maximization problem in the space of continuous *potential* function:

$$\max_{f,g\in\mathcal{C}(\mathcal{X})} \langle f,\, \alpha\rangle + \langle g,\, \beta\rangle + \varepsilon(\langle \alpha\otimes\beta,\, \exp(\frac{f\oplus g - C}{\varepsilon})\rangle - 1).$$

The dual problem (**??**), a regularized version of the dual of (**??**), can be solved by alternated maximization, performing at iteration t

$$f_{t+1}(\cdot) = -T_{C,\varepsilon}(g_t,\beta) \qquad g_{t+1}(\cdot) = -T_{C^\top,\varepsilon}(f_{t+1},\alpha),$$

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

is a *soft C-transform*, and the notation $f_t(\cdot)$ emphasizes the belonging of f_t and g_t to the space of continuous functions. $(f_t)_t$ and $(g_t)_t$ converge in $(\mathcal{C}(\mathcal{X}), \|\cdot\|_{\text{var}})$ to a solution $(f^\star, g^\star)$ of (**??**), 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 $T_{C^\top}(\cdot,\alpha)$ in the space $(\mathcal{C}(\mathcal{X}), \|\cdot\|_{\text{var}})$.

2.2. Estimating OT distances with realizations

Regularized optimal transport is elegantly written in functional spaces but the iterations (**??**) transfers into code only for discrete distributions $\hat\alpha = \sum_{i=1}^n a_i\delta_{x_i}$ and $\hat\beta = \sum_{i=1}^n a_i\delta_{x_i}$. In this case, they correspond to the well-known Sinkhorn-Knopp algorithm for balancing the matrix $\exp(\frac{-C}{\varepsilon})$. The algorithm is run in two phases: a first one, during which the cost matrix is computed (with a cost in $\mathcal{O}(n^2 d)$), and a second one, during which it is balanced (each iteration have a cost in $\mathcal{O}(n^2)$).

Sample complexity results. 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 (**??**). 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 (**??**), 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