Online Sinkhorn: 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.

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 (Arjovsky et al., 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 $\mathcal{X}$. We refer to the monograph of Santambrogio (2015) for a detailed mathematical treatment. 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 once from the input distributions and solving a discrete linear program (LP), due to Kantorovich (1942). This approach is numerically costly and statistically inefficient (Weed & Bach, 2019). The optimisation problem depends on a fixed sampling of points from the data. It is therefore not adapted to machine learning setting where data is resampled continuously (e.g. in GANs), or accessed in an online manner. The goal of this paper is to develop an efficient online method able to estimate OT distances between continuous distributions. We will use a stream of data to refine an approximate optimal transport solution, adapting the celebrated Sinkhorn algorithm to an online setting.

To alleviate both the computational and statistical burdens of OT, it is common to regularize the Kantorovich LP. The most successful approach in this direction is to use an entropic barrier penalty. When dealing with discrete distributions, this yields a problem that 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 (2013). Sinkhorn distances are smooth and amenable to GPU computations, which makes them suitable as a loss function in model training. The Sinkhorn algorithm operates in two distinct phases: draw samples from the distributions and evaluate a pairwise distance matrix in the first phase; balance this distance matrix using Sinkhorn-Knopp iterations in the second phase.

This two-step approach does not estimate the true regularized OT distance, and cannot handle samples provided as a stream, e.g. renewed at each training iteration of an outer algorithm. A cheap fix is to use Sinkhorn over mini-batches (see for instance Genevay et al. (2018) for an application to GANs). Yet this introduces a strong estimation bias, especially in high dimension —see Fatras et al. (2019) for a mathematical analysis. In contrast, we use streams of mini-batches to progressively enrich a consistent representation of the transport plan.

Contributions. Our paper proposes a new take on estimating optimal transport distances between continuous distributions. We make the following contribution

- We introduce an online variant of the Sinkhorn algorithm, that uses streams of samples $(x_t)_t$ and $(y_t)_y$ to enrich a non-parametric functional representation of the dual regularized optimal transport solution.
- We cast online Sinkhorn as an instance of a block-convex stochastic mirror descent algorithm. This

[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.

shows convergence of the algorithm with high probability. We analyse special cases of online Sinkhorn.

- We demonstrate the performance of online Sinkhorn for estimating OT distances between continuous distributions and for accelerating the early phase of discrete Sinkhorn iterations. Comparison with other methods advocates for our original non-parametric representations of OT solutions.

Notations. We denote $\mathcal{C}(\mathcal{X})$ the set of continuous functions over a space $\mathcal{X}$, $\mathcal{M}^+(\mathcal{X})$ and $\mathcal{P}(\mathcal{X})$ the set of positive and probability measures on $\mathcal{X}$, respectively. $\frac{d\mu}{d\alpha}$ denotes the Radon-Nikodym derivative of measure μ with respect to measure α. We write $(i, j]$ the sequence $[i+1, \dots, j]$.

1. Related work

We review recent work on Sinkhorn distances and on the general problem of estimating optimal transport distances.

Sinkhorn properties. Sinkhorn algorithm computes ε-accurate approximations of OT in $O(n^2/\varepsilon^3)$ operations for a number n of samples (Altschuler et al., 2017) (in contrast to the $O(n^3)$ complexity for an exact solution). Moreover, these Sinkhorn distances does not suffers from the curse of dimensionality (Genevay et al., 2019), since the average error using n random samples decays like $O(\varepsilon^{-d/2}/\sqrt{n})$ in dimension d, in sharp contrast with the slow $O(1/n^{1/d})$ error decay of OT (Dudley; Weed & Bach, 2019). These Sinkhorn distances can furthermore be sharpened by entropic debiasing (Feydy et al., 2019). Our work is rather orthogonal to these references, as it focuses on estimating distances between continuous distributions.

Continuous optimal transport. Extending OT computations to arbitrary distributions (possibly having continuous densities) without relying on 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 Wasserstein barycenters).

When dealing with arbitrary continuous densities, which are accessed through a stream of random samples, the challenge is to approximate the (continuous) dual variables of the regularized Kantorovich LP using parametric or non-parametric classes of functions. For application to generative model fitting, one can use deep networks, which leads to an alternative formulation of Generative Adversarial Networks (GANs) (Arjovsky et al., 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. To our knowledge, the only mathematically rigorous algorithm uses reproducing Hilbert space representations of potentials (Genevay et al., 2016). As this construction is generic to all optimisation problems over functions, the convergence is slow. The representations we introduce outperform RKHS representations (§5.3).

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 realizations.

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 distributions $(\alpha, \beta) \in \mathcal{P}(\mathcal{X})^2$ 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 $\mathcal{X} \times \mathcal{X}$):

$$\mathcal{W}_{C,0}(\alpha, \beta) \triangleq \min_{\pi \in \mathcal{P}(\mathcal{X}^2)} \left\{ \langle C, \pi \rangle \ : \ \pi_1 = \alpha, \pi_2 = \beta \right\},$$

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 $\mathcal{W}_{C,0}^{\frac{1}{p}}$ 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 (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

$$\mathcal{W}_{C,\varepsilon}(\alpha, \beta) \triangleq \min_{\substack{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) \\ \pi_1 = \alpha, \pi_2 = \beta}} \langle C, \pi \rangle + \varepsilon \mathrm{KL}(\pi | \alpha \otimes \beta), \quad (1)$$

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). $\mathcal{W}_{C,\varepsilon}$ approximate $\mathcal{W}_C(\alpha, \beta)$ up to an $\varepsilon \log(\varepsilon)$ error (Genevay et al., 2019). In the following, we set ε to 1 without loss of generality, as $\mathcal{W}_{C,\varepsilon} = \varepsilon \mathcal{W}_{C/\varepsilon,1}$, and simply write $\mathcal{W}$. The regularized problem (1) admits a dual form, which is a maximization problem over the space of continuous functions:

$$\max_{(f,g)\in\mathcal{C}(\mathcal{X})^2} \langle f,\, \alpha\rangle + \langle g,\, \beta\rangle - \langle e^{f\oplus g - C},\, \alpha\otimes\beta\rangle + 1, \quad (2)$$

where $\langle f,\, \alpha\rangle \triangleq \int f(x)\mathrm{d}\alpha(x)$ and $(f \oplus g - C)(x, y) \triangleq f(x) + g(y) - C(x, y)$. We write $F_{\alpha,\beta}(f, g)$ the dual objective. Problem (2) can be solved by alternated maximization, which is itself performed in closed form. At iteration t, the updates are simply

$$f_{t+1}(\cdot) = -T_\beta(g_t),\; g_{t+1}(\cdot) = -T_\alpha(f_{t+1}),$$
$$T_\mu(h) \triangleq -\log \int_{y\in\mathcal{X}} \exp(h(y) - C(\cdot, y))\mathrm{d}\mu(y). \quad (3)$$

The operation $h \mapsto T_\mu(h)$ maps a continuous function to another continuous function, and is a smooth approximation of the celebrated C-transform of OT (Santambrogio, 2015). We thus refer to it as a *soft C-transform*. The notation $f_t(\cdot)$ emphasizes the fact that f_t and g_t are *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), where $\|f\|_{\text{var}} \triangleq \max_x f(x) - \min_x f(x)$ is the so-called variation norm. Functions endowed with this norm are only considered up to an additive constant. Global convergence is due to the strict contraction of the operators $T_\beta(\cdot)$ and $T_\alpha(\cdot)$ in the space $(\mathcal{C}(\mathcal{X}), \|\cdot\|_{\text{var}})$ (Lemmens & Nussbaum, 2012).

2.2. Estimating OT distances with realizations

Iterations (3) cannot be implemented when dealing with generic distributions (α, β), because they involve continuous functions $(f_t, g_t)_t$. When the input distribution are discrete (or equivalently that $\mathcal{X}$ is a finite set) then these function can be stored on discrete vectors; the iterations (3) correspond to the celebrated Sinkhorn & Knopp (1967) algorithm, which is often implemented over the scaling variable (e^{f_t}, e^{g_t}). More precisely, with $\alpha = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\beta = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, the Sinkohrn iterations (3) update $u_t \triangleq (e^{f_t(x_i)})_{i=1}^n, v_t \triangleq (e^{g_t(y_i)})_{i=1}^n$ as

$$u_{t+1} = \frac{n}{Kv_t} \quad \text{and} \quad v_{t+1} = \frac{n}{K^\top u_{t+1}}$$

where $K = (e^{-C(x_i,y_i)})_{i,j=1}^n \in \mathbb{R}^{n\times n}$. The algorithm thus operates in two phases: first, the kernel matrix K is computed, with a cost in $O(n^2 d)$, where d is the dimension of $\mathcal{X}$; second, K is balanced, each iteration costing $O(n^2)$. The online Sinkhorn algorithm that we propose mixes these two phases to accelerate convergence (see §5.2).

Consistency and bias. The OT distance $\mathcal{W}_{C,0}(\alpha, \beta)$ and its regularized version $\mathcal{W}_{C,\varepsilon}(\alpha, \beta)$ can be approximated by the (computable) distance between discrete realizations $\hat\alpha = \frac{1}{n}\sum_i \delta_{x_i}$, $\hat\beta = \frac{1}{n}\sum_i \delta_{y_i}$, where $(x_i)_i$ and $(y_i)_i$ are i.i.d samples from α and β. Consistency holds, as $\mathcal{W}_{C,\varepsilon}(\hat\alpha_n, \hat\beta_n) \to \mathcal{W}_{C,\varepsilon}(\alpha, \beta)$. Although this is a reassuring result, the sample complexity of transport in high dimensions with low regularization remains high (see §1). For computational reasons, we cannot choose n to be much more than 10^5. We may wonder wether we can improve the estimation of $\mathcal{W}_{C,\varepsilon}(\alpha, \beta)$ using several sets of samples $(\hat\alpha_t)_t$ and $(\hat\beta_t)_t$. Those should be of reasonable size to allow Sinkhorn estimation, and may for example come from a temporal stream. Genevay et al. (2018) propose to use a Monte-Carlo estimate $\hat{\mathcal{W}}(\alpha, \beta) = \frac{1}{T}\sum_{t=1}^T \mathcal{W}(\hat\alpha_t, \beta_t)$. However, this yields a biased estimation as the distance $\mathcal{W}(\alpha, \beta)$ differs from its expectation under sampling $\mathbb{E}_{\hat\alpha\sim\alpha,\hat\beta\sim\beta}[\mathcal{W}(\hat\alpha, \hat\beta)]$. To address this issue, our algorithm computes a consistent estimation of the Sinkhorn cost using a stream of samples.

Similarly, estimating the Sinkhorn potentials through sampling is biased, as the true optimal potential $f^\star = f^\star(\alpha, \beta)$ is not the expected optimal potential under sampling $\mathbb{E}_{\hat\alpha_n\sim\alpha,\hat\beta_n\sim\beta}[f^\star(\hat\alpha_n, \hat\beta_n)]$. Our algorithm estimates the true continuous potential functions (up to a constant) and overcomes the sampling bias.

3. OT distances from sample streams

We introduce an online adaptation of the Sinkhorn algorithm, the major contribution of this paper. We construct an estimator of $f^\star$, $g^\star$ and $\mathcal{W}(\alpha, \beta)$ using successive sets of samples $(\hat\alpha_t)_t = \frac{1}{n}\sum_{i=n_t+1}^{n_{t+1}} \delta_{x_i}$ and $(\hat\beta_t)_t$, where we set $n_t = nt$. $(\hat\alpha_t)_t$ and $(\hat\beta_t)_t$ may be seen as mini-batches of size n within a training procedure, or as a temporal stream of samples. Our estimator progressively enriches a representation of the solution of (2), that may be arbitrary complex.

We detail an intuitive construction of our algorithm in §3.1, formalize it in §3.2 before casting it as a block-convex stochastic mirror descent in 3.3.

3.1. Online Sinkhorn iterations

From (2), along the continuous (and untractable) Sinkhorn optimisation trajectory $(\bar f_t, \bar g_t)_t$, the potential $\bar f_t$ is always the negative logarithm of an infinite mixture of kernel functions $\kappa_y(x) \triangleq \exp(-C(\cdot, y))$:

$$\exp(-\bar f_t(\cdot)) = \int_{\boldsymbol{y}\in\mathcal{X}} \exp(\bar g_t(y))\kappa_y \mathrm{d}\beta(y),$$

and similarly for $\bar g_t$. Our algorithm constructs a sequence of non-parametric potentials $(f_t, g_t)_t$ that behaves as $(\bar f_t, \bar g_t)$. The strong structural property of the continuous potentials

suggests to express $\exp(-f_t)$ as a finite mixture of kernel functions. That is, f_t and g_t are continuous functions constructed respectively from the weights $(p_i, q_i)_{i\leqslant n_t}$ and positions $(x_i, y_i)_i \subset \mathcal{X}$ as

$$f_t(\cdot) = -\log \sum_{i=1}^{n_t} \exp(q_i - C(\cdot, y_i)), \qquad (4)$$
$$g_t(\cdot) = -\log \sum_{i=1}^{n_t} \exp(p_i - C(x_i, \cdot)).$$

Randomized Sinkhorn. Provided with fresh samples $(x_i, y_i)_{n_t<i\leqslant n_{t+1}}$, corresponding to empirical measures $\hat{\alpha}_t$ and $\hat{\beta}_t$, a naive approach would update the potentials using a noisy soft C-transform:

$$f_{t+1} = T_{\hat{\beta}_t}(g_t), \qquad g_{t+1} = T_{\hat{\alpha}_t}(f_{t+1}), \qquad (5)$$

which is equivalent to setting all $(q_i)_{i\leqslant n_t}$ to 0, and assigning each weight q_i to $g_t(y_i) - \log(n)$ for $n_t < i \leqslant n_{t+1}$ and similarly for p_i. The variance of the updates (5) does not decay through the iteration, hence *random Sinkhorn* algorithm does nos converge. However, thanks to the contraction of the random operator $T_{\hat{\beta}_t}(\cdot)$ and $T_{\hat{\alpha}_t}(\cdot)$, Prop. 2 shows that that the Markov chain $(f_t, g_t)_t$ that it defines converges towards a stationary distribution independent of initialization.

Online Sinkhorn. To ensure convergence towards the potentials $(f^\star, g^\star)$ (up to a constant factor), we must therefore take more cautious steps—in other words, we cannot afford to forget past iterates to obtain a consistent estimation of potentials. We introduce a learning rate in Sinkhorn iterations, that averages the past representations and the newly computed noisy C-transforms.

$$\exp(-f_{t+1}) \triangleq (1-\eta_t)\exp(-f_t) + \eta_t \exp(-T_{\hat{\beta}_t}(g_t)), \qquad (6)$$

and similarly for g_t. Performing the averaging over the space of inverse scalings $(e^{-\hat{f}_t}, e^{-\hat{g}_t})$ yields simple updates for the weights $(p_i, q_i)_i$, and is crucial for our theoretical convergence analysis. In essence, the weights of past samples are reduced by a constant factor, while new weights are computed from the evaluation of $f_t(\cdot), g_t(\cdot)$ at random new points $(x_i, y_i)_i$. Note that we perform *simultaneous updates* of f_t and g_t, which is important for the convergence analysis.

Estimating Sinkhorn distance. The iterations (6) allow to estimate potential functions up to a constant. As explained in §2.2, this estimation is sufficient for most applications aiming at minimizing a Sinkhorn loss, as it only requires the spatial derivatives of the potentials. If required, it is however possible to estimate the Sinkhorn distance

Algorithm 1 Online Sinkhorn potentials

Input: Distribution α and β, learning weights $(\eta_t)_t$
Set $p_i = q_i = 0$ for $i \in (0, n_t]$
for $t = 0, \dots, T-1$ **do**
  for $i \in (0, n_t]$ **do**
    $q_i \leftarrow q_i + \log(1-\eta_t)$, $p_i \leftarrow p_i + \log(1-\eta_t)$,
  Sample $(x_i)_{(n_t, n_{t+1}]} \sim \alpha$, $(y_j)_{(n_t, n_{t+1}]} \sim \beta$
  for $i \in (n_t, n_{t+1}]$ **do**
    $q_i \leftarrow \log(\eta_t) - \log \frac{1}{n_t} \sum_{j=1}^{n_t} \exp(p_j - C(x_j, y_i))$
    $p_i \leftarrow \log(\eta_t) - \log \frac{1}{n_t} \sum_{j=1}^{n_t} \exp(q_j - C(x_i, y_j))$
  Optional: refit all $q_i = g_t(y_i) - \log(n_{t+1})$
    $p_i = f_t(x_i) - \log(n_{t+1})$
  Save $(q_i, p_i, x_i, y_i)_{(n_t, n_{t+1}]}$
Returns $f_T : (q_i, y_i)_{(0, n_T]}$ and $g_T : (p_i, x_i)_{(0, n_T]}$

using our method, by performing a final soft C-transform, using $\mathcal{O}(n_T^2)$ operations:

$$\mathcal{W}_t = \frac{1}{2}\Big(\langle \bar{\alpha}_T,\, f_T + T_{\bar{\alpha}_T}(g_T)\rangle + \langle \bar{\beta}_T,\, g_T + T_{\bar{\alpha}_T}(f_T)\rangle\Big), \qquad (7)$$

where $\bar{\alpha}_t \triangleq \frac{1}{n_{t+1}} \sum_{i=1}^{n_{t+1}} \delta_{x_i}$ and $\bar{\beta}_t$ gather previously observed samples.

3.2. Algorithm, complexity and refinements

The pseudo-code of online Sinkhorn is detailed in Alg. 1. We perform the updates for q_i and p_i in log-space, for numerical stability reasons. Each iteration has complexity $\mathcal{O}(t\, n^2)$, due to the evaluation of the distances $C(x_i, y_i)$ for all $(x_i)_{(0,n_t]}$, $(y_i)_{(n_t, n_{t+1}]}$ and to the computation of the soft C-transforms. Online Sinkhorn estimates and records a distance matrix $(C(x_i, y_j))_{i,j}$ on the fly, in parallel to the updates of the potentials f_t and g_t. In total, its computation cost after drawing n_t samples is $\mathcal{O}(n_t^2)$, and its memory cost is $\mathcal{O}(n_t)$. We propose some heuristics to accelerate convergence and alleviate memory and computational cost.

Fully-corrective scheme. The potentials f_t and g_t may be improved by refitting the weights $(p_i)_{(0,n_t]}$, $(q_j)_{(0,n_t]}$ based on all previously seen samples. This amounts to replace, after iteration t, for all $i \in (0, n_{t+1}]$,

$$q_i \leftarrow g_t(y_i) - \log(n_{t+1}) = -\log \frac{1}{n_{t+1}} \sum_{j=1}^{n_{t+1}} e^{p_j - C(x_j, y_i)},$$

and similarly for each p_i. This corresponds to performing one step of the discrete Sinkhorn algorithm with distributions $\bar{\alpha}_t$ and $\bar{\beta}_t$. This increases the dual cost $F_{\bar{\alpha},\bar{\beta}}(f_t, g_t)$, and "on average", the energy $F_{\alpha,\beta}$. This reweighted scheme (akin to the fully-corrective Frank-Wolfe scheme from Lacoste-Julien & Jaggi (2015)) has a cost in $\mathcal{O}(n_t^2)$. In

practice, it can be used only every k iterations, with k increasing with t. We study a combination of partial and full updates in §5.2.

Memory compression. The memory requirement in $\mathcal{O}(n_t)$ is an avoidable limitation of the algorithm, as the optimal potentials $(f^\star, g^\star)$ do not admit a parametric representation in general. However, we may compress the representations (q_j, y_j) and $(x_i, p_i)_i$ using k-means clustering over M centroids. The sampled points $(x_i)_i$ and $(y_j)_j$ are attached to centroids $(X_I)_{I\in(0,M_t]}$ and $(Y_J)_{J\in(0,M_t]}$. For all $I \in (0, M_t]$, we set weights and potentials as

$$Q_J \leftarrow -\log \sum_{j,\ y_j \text{ closest} \atop \text{to } \bar{Y}_J} \exp(-q_j),$$

$$f_t(\cdot) \leftarrow -\log \sum_{J=1}^{M_t} \exp(Q_J - C(\cdot, \bar{Y}_J)),$$

and similarly for $(p_I)_I$ and g_t. Once again, this operation should be made once every k iterations. M_t can for instance be set constant after linearly increasing in a first stage. This heuristic is important for applications but requires significant engineering: we leave it for future work.

Out-of-loop averaging, finite samples. Optionally, we may also compute out-of-loop averages of potentials

$$\exp(-\bar{f}_{t+1}) = (1-\gamma_t)\exp(-\bar{f}_t) + \gamma_t \exp(-\hat{f}_t),$$
$$\exp(-\bar{g}_{t+1}) = (1-\gamma_t)\exp(-\bar{g}_t) + \gamma_t \exp(-\hat{g}_t),$$

to further reduce the estimation variance. We show in §5 that this averaging is efficient in practice. Finally, we note that our algorithm applies on both continuous or discrete distributions. When α and β are discrete distributions of size N, we can store p and q as fixed-size vectors of size N, and subsample mini-batches of size $n < N$. The resulting algorithm is a *subsampled* Sinkhorn algorithm for histograms, which is detailed in the appendix for completeness. We show in §5 that it is useful to accelerate the first phase of the Sinkhorn algorithm.

3.3. Stochastic mirror descent interpretation

This online Sinkhorn can be recast as a stochastic mirror descent algorithm, which enables a convergence analysis. This equivalence is obtained by applying a change of variable in (1), defining

$$\mu \triangleq \alpha \exp(f) \quad \text{and} \quad \nu \triangleq \beta \exp(g). \tag{8}$$

The dual problem (2) rewrites as a minimisation problem over positive measures on $\mathcal{X}$ and $\mathcal{Y}$:

$$-\min_{(\mu,\nu)\in\mathcal{M}^+(\mathcal{X})^2} \mathrm{KL}(\alpha|\mu) + \mathrm{KL}(\beta|\nu) + \langle \mu\otimes\nu,\, e^{-C}\rangle - 1, \tag{9}$$

where the function $\mathrm{KL} : \mathcal{P}(\mathcal{X}) \times \mathcal{M}^+(\mathcal{X}) \triangleq \langle \alpha, \log \frac{\mathrm{d}\alpha}{\mathrm{d}\mu}\rangle$ is the Kullback-Leibler divergence between α and μ. This objective is block convex in μ, ν, but not jointly convex. As we now detail, this problem can be solved using a stochastic mirror descent (Beck & Teboulle, 2003), applied here over the Banach space of Radon measures on $\mathcal{X}$, equipped with the total variation norm.

Mirror maps and gradient. For this, we define the (convex) distance generating function $\mathcal{M}^+(\mathcal{X})^2 \to \mathbb{R}$:

$$\omega(\mu,\nu) \triangleq \mathrm{KL}(\alpha|\mu) + \mathrm{KL}(\beta|\nu).$$

The gradient of this function and of its Fenchel conjugate $\omega^\star : \mathcal{C}(\mathcal{X})^2 \to \mathbb{R}$ yields two *mirror maps*. For all $(\mu,\nu) \in \mathcal{M}^+(\mathcal{X})^2$, $(\varrho,\varphi) \in \mathcal{C}(\mathcal{X})^2$, $\varrho < 0, \varphi < 0$,

$$\nabla\omega(\mu,\nu) = (-\frac{\mathrm{d}\alpha}{\mathrm{d}\mu}, -\frac{\mathrm{d}\beta}{\mathrm{d}\nu}) \qquad \nabla\omega^\star(\varrho,\varphi) = (-\frac{\alpha}{\varrho}, -\frac{\beta}{\varphi}).$$

The gradient $\nabla F(\mu,\nu)$ of the objective F appearing in (9) is a continuous function

$$\nabla_\mu F(\mu,\nu) = -\frac{1}{\frac{\mathrm{d}\mu}{\mathrm{d}\alpha}} + \int_{y\in\mathcal{X}} \frac{\mathrm{d}\nu}{\mathrm{d}\beta}(y)\exp(-C(\cdot,y))\mathrm{d}\beta(y)$$

and similarly for $\nabla_\nu F$.

Stochastic mirror descent. To define stochastic mirror descent iterations, we may replace integration over β is by an integration over a sampled measure $\hat{\beta}$. This in turn defines an *unbiased gradient estimate* $\tilde{\nabla}F$ of ∇F, which has bounded second order moments. This absence of bias is crucial to prove convergence of SMD with high probability. Using the mirror maps and the stochastic estimation of the gradient, one has the following equivalence result, whose proofs stems from direct computations.

Proposition 1. *The stochastic mirror descent iterations*

$$(\mu_t,\nu_t) = \nabla\omega^\star\Big(\nabla\omega(\mu_t,\nu_t) - \eta_t\tilde{\nabla}F(\mu_t,\nu_t)\Big)$$

are equal to the updates (6) *under the change of variable* (8).

Interpretation. It is important to realize that μ_t and ν_t does not need to be stored in memory. Instead, their associated potentials f_t and g_t are parametrized as (4). In particular, μ_t and ν_t remains absolutely continuous with respect to α and β respectively, so that the Kullbach-Leibler divergence terms are always finite. Note that the mirror descent we consider operates in an infinite-dimensional space, which is valid under mild conditions (Hsieh et al., 2018).

Finally, we mention that when computing exact gradient (in the absence of noise) and when using constant step-size of $\eta_t = 1$, the algorithm matches exactly Sinkhorn iterations with simultaneous updates of the dual variable. This

provides a novel interpretation on the Sinkhorn algorithm, that differs from the usual Bregman projection (Benamou et al., 2015), and the related understanding of Sinkhorn as a constant step-size mirror descent on the primal objective (Mishchenko, 2019).

4. Convergence analysis

We give three kind of convergence analysis: (i) a stationary distribution convergence property for the random Sinkhorn algorithm ; (ii) a global convergence property for the online Sinkhorn algorithm without noise ; (iii) a high-probability convergence result for the full online Sinkhorn algorithm.

Randomized Sinkhorn. We first state a result concerning the randomized Sinkhorn algorithm (6), which corresponds to Alg. 1 with step-size $\eta_t = 1$.

Proposition 2. *The random Sinkhorn algorithm (6) yields a time-homogeneous Markov chain $(f_t, g_t)_t$ which is $(\hat{\alpha}_s, \hat{\beta}_s)_{s\leqslant t}$ measurable, and converges in law towards a stationary distribution $(F_\infty, G_\infty) \in \mathcal{P}(\mathcal{C}(\mathcal{X})^2)$ independent of the initialization point (f_0, g_0).*

This result follows from Diaconis & Freedman (1999) convergence theorem on iterated random functions which are contracting on average. We simply use the fact that $T_{\hat{\beta}}(\cdot)$ and $T_{\hat{\alpha}}(\cdot)$ are *always* contracting, independent of the distributions $\hat{\alpha}$ and $\hat{\beta}$, for the variational norm $\|\cdot\|_{\text{var}}$.

Note that using the law of large number for Markov chains (Breiman, 1960), the out-of-loop averages $\exp(-\bar{f}_t)$ for converge to $\mathbb{E}[\exp(-F_\infty)] \in \mathcal{C}(\mathcal{X})$ for $\gamma_t = \frac{1}{t}$. This expectation verifies the following fixed point equations

$$\mathbb{E}[\exp(-F_\infty)] = \langle \beta,\, \mathbb{E}[\exp(G_\infty)]\exp(-C)\rangle$$
$$\mathbb{E}[\exp(-G_\infty)] = \langle \alpha,\, \mathbb{E}[\exp(F_\infty)]\exp(-C)\rangle.$$

These fixed point equations are close to the Sinkhorn fixed point equations, and get closer as ε increases, since $\varepsilon\mathbb{E}[\exp(\pm F_\infty/\varepsilon)] \to \mathbb{E}[F_\infty]$ as $\varepsilon \to \infty$. Running the random Sinkhorn algorithm with out-of-loop averaging fails to provide exactly the dual solution. However, it defines an approximate solution of the original problem whose accuracy depends on ε. We leave the quantification of this approximation for future work.

Noise-free online Sinkhorn. Variance reduction is therefore necessary, to ensure that the limit stationary distribution is deterministic. The following proposition shows that the modified "slowed-down" online Sinkhorn algorithm converges in the absence of noise.

Proposition 3. *We suppose that $\hat{\alpha}_t = \alpha$, $\hat{\beta}_t = \beta$ for all t. Then the updates (6) yields a sequence $(f_t, g_t)_t$ such that*

$$\|f_t - f^\star\|_{var} + \|g_t - g^\star\|_{var} \to 0$$
$$\frac{1}{2}\langle \alpha,\, f_t + T_\alpha(g_t)\rangle + \langle \beta,\, g_t + T_\beta(f_t)\rangle \to \mathcal{W}(\alpha, \beta).$$

Note that, due to the fact that we perform simultaneous updates, we only obtain the convergence of $f_t \to f^\star + A$, and $g_t \to g^\star$, where $f^\star$ and $g^\star$ are solutions of (1) and A is a constant depending on initialization. This is only a small caveat, as we can average the potentials and their soft C-transform as in (7) to remove the offset A. This is not necessary when using the Sinkhorn distance as a loss for training purposes, e.g. for generative modeling or barycenter estimation as in Staib et al. (2017). Backpropagation through the Sinkhorn distance indeed relies only on the gradients of the potentials $\nabla_x f^\star(\cdot)$, $\nabla_y g^\star(\cdot)$ (e.g. Cuturi & Peyré, 2018).

Online Sinkhorn. Finally, we state a result on the complete online Sinkhorn algorithm, which converges with high probability.

Proposition 4. *Assume that $\|f_0 - f^\star\|_{var} \leqslant 1$ and $\|g_0 - g^\star\|_{var} \leqslant 1$, $\sum_t \eta_t = \infty$ and $\sum_t \eta_t^2 < \infty$ and fix a level of confidence δ. Then, with probability $1-\delta$, provided that η_t is small enough,*

$$\|f_t - f^\star\|_{var} + \|g_t - g^\star\|_{var} \to 0$$
$$\mathcal{W}_t = \frac{1}{2}\langle \alpha,\, f_t + T_\alpha(g_t)\rangle + \langle \beta,\, g_t + T_\beta(f_t)\rangle \to \mathcal{W}(\alpha, \beta).$$

This result relies on Theorem 5.2 from Zhou et al. (2017), that establishes convergence with high probability of stochastic mirror descent applied to non-convex but locally coherent objectives. In particular, the proof of this results does not rely on the finite dimension of the ambiant space.

5. Experiments

We have introduced and stated convergence results on the online Sinkhorn algorithm. These convergence results are non-quantitative and therefore require an extensive experiment validation. Our experiments are three-fold: first, we show that online Sinkhorn correctly estimates the solutions of (1) and the Sinkhorn distance, overcoming the bias due to the fixed a priori sampling of the regular Sinkhorn algorithm. Then, we show how online Sinkhorn accelerates the Sinkhorn algorithm, by progressively estimating sketches of the dual potentials, in parallel to the computation of the distance matrix. Finally, we show how online Sinkhorn allows one to estimate accurately the geometry of the dual, significantly improving the result using SGD with RKHS expansions (Genevay et al., 2016). Numpy and Pytorch code are provided with this manuscript.

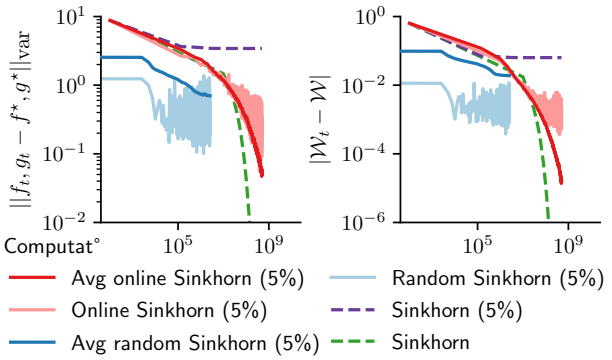

Figure 1. Comparison of online, random and fixed sampling Sinkhorn performances. Online Sinkhorn overcomes the bias of sampling—especially with out-of-loop averaging. Random Sinkhorn gives fast estimations, whose variance does not decrease.

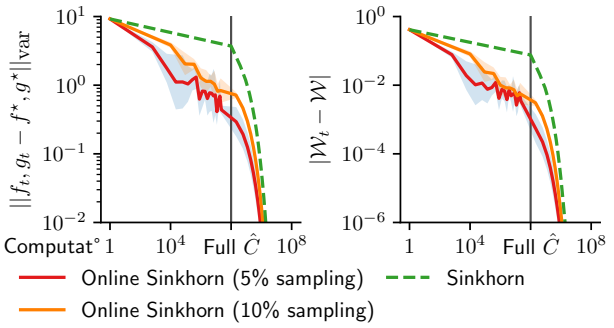

Figure 2. Using online Sinkhorn during the initial computation of the cost matrix accelerates the Sinkhorn algorithm: it provides good estimates of the potentials f and g to warm start the full Sinkhorn algorithm. Curves averaged over 5 runs.

5.1. Consistent estimation of Sinkhorn distances

We first consider a discrete distribution (α, β), to be able to compute the reference distance $\mathcal{W} = \mathcal{W}(\alpha, \beta)$ and the optimal potentials $f^\star$, $g^\star$, using Sinkhorn algorithm. The goal here is not to perform better than the Sinkhorn algorithm in the long run. Indeed, the constraints of online Sinkhorn yields unnecessary slow-downs when dealing with small discrete distributions. Rather, our purpose is to illustrate the consistence of online Sinkhorn. We choose α and β to be two discrete 1-D distributions, $\mathcal{X} = \mathbb{R}$, sampled from the continuous densities displayed in Fig. 3. We set $\varepsilon = 10^{-2} \max_{x,y} C(x, y)$, where we use the squared Euclidean loss (regularized $\mathcal{W}_2$ setting)—the distributions α and β have bounded support. We use $\eta_t = \frac{1}{\sqrt{t}}$ for online Sinkhorn, in all experiments. We compare the performance of Sinkhorn, online Sinkhorn and random Sinkhorn, measuring $\|f - f^\star\|_{\text{var}} + \|g - g^\star\|_{\text{var}}$ and the absolute error $|\mathcal{W}_t - \mathcal{W}|$ versus the number of computations performed—the evaluation of $C(x_i, y_i)$ and the computation of each addition in the C-transform being considered as elementary computation units. We further report the performance of using out-of-loop averaging with $\gamma_t = \frac{1}{\sqrt{t}}$.

Results. We report convergence curves in Fig. 1. Compared to the subsampled Sinkhorn algorithm that computes a biased estimate of the distance $\mathcal{W}$ (purple), the online Sinkhorn algorithm successfully estimates the distance and the associated potentials, despite performing only partial C-transforms (red). Random Sinkhorn (blue) finds a decent estimation of the distance and potentials, with fewer computations than the full Sinkhorn algorithm, but fails to converge. Averaging the random Sinkhorn iterations finds a biased estimation. The vanilla online Sinkhorn converges towards the true value, albeit with a rather high iterate variance (note that this variance does reduce—this is a log-log plot). Remarkably, the out-of-loop averaging of online Sinkhorn enjoys much better converging property—we confirmed this finding on many synthetic problems. It is surprising that an averaging mechanism brings speed-up in a non-convex setting—we attribute this to the convexity of the original problem, although this should be further investigated.

5.2. Accelerating the first Sinkhorn iteration

The discrete Sinkhorn algorithm requires to compute a full matrix $\hat{C} \triangleq (C(x_i, y_i))_{i,j}$ of size $N \times N$, prior to estimating the potentials f_1 and g_1 by a first C-transform. In contrast, online Sinkhorn can progressively computes this matrix while computing first sketches of the potentials. We therefore assess the performance of the following *online+full Sinkhorn* algorithm in a discrete setting: online Sinkhorn is run with batch-size n during the first iterations, until observing each sample of $[1, N]$, i.e. until the cost matrix C is completely evaluated. At this point (iteration t), online Sinkhorn provides the estimates f_t, g_t. From then, the algorithm only performs full Sinkhorn updates.

Results. We report convergence curves in Fig. 2. The proposed scheme indeed provides an improvement upon Sinkhorn algorithm. After N^2 computation (the cost of estimating the full matrix $\hat{C}$), both the function value and distance to optimum are lower using our scheme: the full Sinkhorn algorithm then operates from a good initialization for potentials. Computing those cost approximately as much as estimating the matrix $\hat{C}$ in dimension 1. The *online+full* Sinkhorn algorithm then maintain an advantage over the full Sinkhorn algorithm over time. Note that the cost of estimating initial potentials becomes negligible as the dimension increase—the cost of computing $\hat{C}$ dominates. This strongly advocates for using an online scheme as a warm-up for regularized OT estimation. We note that using smaller batch-size n may lead to higher speed-up (here, $n = 50$ performs better than $n = 100$). There is an

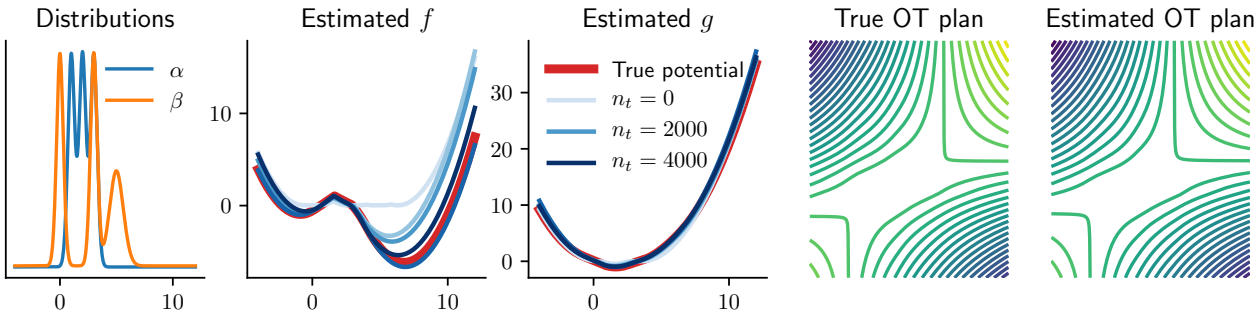

Figure 3. Representation of the convergence path of online Sinkhorn: the blue curves represents the estimated potentials (continuous functions) at different stages of the algorithm. The estimated plan π_t is very quickly accurate, while the shape of the potentials match nearly perfectly the true potentials (estimated on a grid $N = 5000$). $\varepsilon = 10^{-2} \max \hat{C}$.

optimal n. The speed gain decreases with ε, but remains significant even for $\varepsilon = 10^{-4} \max \hat{C}$. We add that using a sampling-without-replacement scheme brings an additional speed-up. Out-of-loop averaging is also beneficial. We refer to the appendix for additional figures.

5.3. Continuous potential estimation

Finally, we measure the performance of our algorithm in a truly continuous setting, where α and β are 1-D parametric distributions (Gaussian mixtures) from which we sample. In the absence of reference $\mathcal{W}$ (which cannot be accurately computed without a method akin to ours), we monitor the trajectories of the potentials, and compare them to the Sinkhorn potentials for realization of α and β of size $n = 2000$. We also monitor the estimated transportation plan $\hat{\pi}_t = (\alpha \otimes \beta) \exp(\frac{f \oplus g - C}{\varepsilon}) \in \mathcal{M}^+(\mathcal{X})^2$. We run the experiments with $n_T = 5000$.

Results. We show the convergence trajectories of the potentials in Fig. 3. Online Sinkhorn refines the potentials $(f_t, g_t)_t$ until convergence. The fact that our method uses an adapted potential parametrization (4) allows the iterates to quickly identify the correct shape of the optimum. The final plan is undistinguishable from the true transportation plan. Quantitative values (distance to true potentials, error in Sinkhorn distance estimation) converge as in Fig. 1.

Comparison to concurrent approaches. Finally, we compare online Sinkhorn to constructing representations of Sinkhorn potentials using universal RKHS (Genevay et al., 2016). This competing approach sets $f_t(\cdot) = \sum_{i=1}^{n_t} \alpha_t \kappa(\cdot, x_i)$ (and similarly for g_t), where κ is a reproducing kernel (typically a Gaussian). This differs significantly from the representations that we propose, for which $\exp(-f_t)$, and not f_t, is expressed as a kernel mixture. With RKHS representations of potentials, the dual problem (3) can be solved using stochastic gradient descent, with theoretical convergence guarantees. As advocated by the authors,

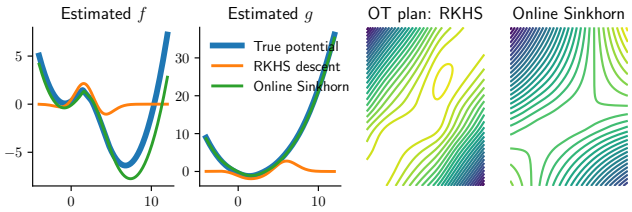

Figure 4. Comparing online Sinkhorn with SGD over a RKHS representation of the potential (Genevay et al., 2016), with best bandwidth parameter. Online Sinkhorn finds more accurate functional representations of potentials, thanks to its more appropriate parametrization. $\varepsilon = 10^{-1} \max \hat{C}$.

we run a grid search over the bandwidth parameter σ of the Gaussian kernel to select the best performing runs. We set $n_T = 50000$, and $\varepsilon = 10^{-1} \max C$. We could not successfully use the RKHS method for lower ε.

We compare the final potentials and associated transportation plans in Fig. 4. Our method estimates potentials with much less errors, especially in areas where the mass of α and β is low. The computational complexity of both algorithm is comparable. Online Sinkhorn does not require to set any hyperparameters, whereas we observed that SGD in RKHS is very sensitive to bandwidth selection.

6. Conclusion

In this article, we have extended the classical Sinkhorn algorithm to cope with streaming samples. The resulting online algorithm computes a non-parametric expansion of the inverse scaling variables using kernel functions. In contrast with previous attempts to compute OT between continuous densities, these kernel expansions fit perfectly the structure of the entropic regularization, which is key to the practical efficiently of our method. We have drawn links between regularized OT ands non-convex mirror descent methods. This in turn opens promising avenues to study convergence rates of continuous variants of Sinkhorn's iterations.

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*, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial network. *Proc ICML'17*, 60(4): 941–965, 2017.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Breiman, L. The strong law of large numbers for a class of markov chains. 31(3):801–803, 1960.

Canas, G. and Rosasco, L. Learning probability measures with respect to optimal transport metrics. In *Advances in Neural Information Processing Systems*, 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*, 2013.

Cuturi, M. and Peyré, G. Semidual regularized optimal transport. *SIAM Review*, 60(4):941–965, 2018.

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.

Diaconis, P. and Freedman, D. Iterated random functions. *SIAM Review*, 41(1):45–76, 1999.

Dudley, R. M. The speed of mean Glivenko-Cantelli convergence. 40(1):40–50.

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 *International Conference on Artificial Intelligence and Statistics*, 2019.

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

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systens*, pp. 3432–3440, 2016.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617, 2018.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. Sample complexity of sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Hsieh, Y.-P., Liu, C., and Cevher, V. Finding mixed nash equilibria of generative adversarial networks. *arXiv preprint arXiv:1811.02002*, 2018.

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

Lacoste-Julien, S. and Jaggi, M. On the global linear convergence of frank-wolfe optimization variants. In *Advances in Neural Information Processing Systems*, 2015.

Lemmens, B. and Nussbaum, R. *Nonlinear Perron–Frobenius Theory*. Cambridge University Press, 2012.

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

Mishchenko, K. Sinkhorn algorithm as a special case of stochastic mirror descent. *arXiv preprint arXiv:1909.06918*, 2019.

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. *International Conference on Learning Representations*, 2018.

Sinkhorn, R. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 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*, 2017.

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

Zhou, Z., Mertikopoulos, P., Bambos, N., Boyd, S., and Glynn, P. On the convergence of mirror descent beyond stochastic convex programming. *arXiv preprint arXiv:1706.05681*, 2017.