Online Sinkhorn: Optimal Transport distances from sample streams

Arthur Mensch, Gabriel Peyré

Online Sinkhorn: Optimal Transport distances from sample streams

Arthur Mensch
ENS, PSL University
Paris, France

Gabriel Peyré
ENS, PSL University
Paris, France

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 provide a theoretical analysis of the convergence of the online Sinkhorn algorithm, showing a nearly-$\mathcal{O}(\frac{1}{n})$ asymptotic sample complexity for the iterate sequence. We validate our method on synthetic 1D to 10D data and on real 3D shape data.

Optimal transport (OT) distances are fundamental in statistical learning, both as a tool for analyzing the convergence of various algorithms (Canas and Rosasco, 2012; Dalalyan and 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 distance over data points living in a space $\mathcal{X}$ into a distance on the space $\mathcal{P}(\mathcal{X})$ of probability distributions over the space $\mathcal{X}$. This distance has many favorable geometrical properties. In particular it allows one to compare distributions having disjoint supports. Computing OT distances 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 and Bach, 2019). Furthermore, the optimisation problem depends on a fixed sampling of points from the data. It is therefore not adapted to machine learning settings where data is resampled continuously (e.g. in GANs), or accessed in an online manner. In this paper, we develop an efficient online method able to estimate OT distances between continuous distributions. It uses a stream of data to refine an approximate OT solution, adapting the regularized OT approach 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 and Knopp, 1967). This approach was pushed forward for ML applications by Cuturi (2013). Sinkhorn distances are smooth and amenable to GPU computations, which make them suitable as a loss function in model training (Frogner et al., 2015; Mensch et al., 2019). 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 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, Peyré, et al. (2018) for an application to generative modelling). 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 contributions:

- We introduce an online variant of the Sinkhorn algorithm, that relies on streams of samples to enrich a non-parametric functional representation of the dual regularized OT solution.
- We establish the almost sure convergence of online Sinkhorn and derive asymptotic convergence rates (Proposition 3 and 4). We provide convergence results for variants.
- We demonstrate the performance of online Sinkhorn for estimating OT distances between continuous distributions and for accelerating the early phase of discrete Sinkhorn iterations.

Notations. We denote $\mathcal{C}(\mathcal{X})$ $[\mathcal{C}_+(\mathcal{X})]$ the set of [strictly positive] continuous functions over a metric space $\mathcal{X}$, $\mathcal{M}^+(\mathcal{X})$ and $\mathcal{P}(\mathcal{X})$ the set of positive and probability measures on $\mathcal{X}$, respectively.

1 Related work

Sinkhorn properties. The Sinkhorn algorithm computes ε-accurate approximations of OT in $O(n^2/\varepsilon^3)$ operations for n samples (Altschuler et al., 2017) (in contrast with the $\mathcal{O}(n^3)$ complexity of exact OT Goldberg and Tarjan, 1989). Moreover, Sinkhorn distances suffer less from the curse of dimensionality (Genevay, Chizat, et al., 2019), since the average error using n samples decays like $\mathcal{O}(\varepsilon^{-d/2}/\sqrt{n})$ in dimension d, in contrast with the slow $\mathcal{O}(1/n^{1/d})$

error decay of OT (Dudley, 1969; Weed and Bach, 2019). Sinkhorn distances can further be sharpened by entropic debiasing (Feydy et al., 2019). Our work is orthogonal, as we focus 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 and Peyré, 2018), and can also be coupled with stochastic optimization methods (Genevay, Cuturi, et al., 2016) to tackle high=dimensional problems (see Staib et al., 2017 for an extension to Wasserstein barycenters). When dealing with arbitrary continuous densities, that 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 represents potentials in reproducing Hilbert space (Genevay, Cuturi, et al., 2016). This approach is generic and does not leverage the specific structure of the OT problem, so that in practice its convergence is slow. We show in Section §5.1 that online Sinkhorn finds better potential estimates than SGD on RKHS representations.

Stochastic approximation (SA). Our approach may be seen as SA (Robbins and Monro, 1951) for finding the roots of an operator in a non-Hilbertian functional space. Alber et al., 2012 studies SA for solving fixed-points that are contractant in Hilbert spaces. Online Sinkhorn convergence relies on the contractivity of a certain operator in a non-Hilbertian metric, and requires a specific analysis. As both are SA instances, the online Sinkhorn algorithm resembles stochastic EM (Celeux and Diebolt, 1992), though it cannot be interpreted as such.

2 Background: optimal transport distances

We first recall the definition of optimal transport distances between arbitrary distributions (i.e. not necessarily discrete), then review how these are estimated using a finite number of 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 $(x, y) \in \mathcal{X}^2 \to C(x, y) \in \mathbb{R}$ 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. We refer to Santambrogio, 2015 for a review on OT.

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 strict convexity. The so-called Sinkhorn cost is then

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

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). $\mathcal{W}_{C,\varepsilon}$ approximates $\mathcal{W}_{C,0}(\alpha, \beta)$ up to an $\varepsilon \log(\varepsilon)$ error (Genevay, Chizat, 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 $\mathcal{W}$. (1) admits a dual form, which is a maximization problem over the space of continuous functions:

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

where $\langle f, \alpha \rangle \triangleq \int f(x) \mathrm{d}\alpha(x)$ and $(f \oplus g - C)(x, y) \triangleq f(x) + g(y) - C(x, y)$. Problem (2) can be solved by closed-form alternated maximization, which corresponds to Sinkhorn's algorithm. At iteration t, the updates are simply

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

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*. Note that

we consider *simultaneous* updates of f_t and g_t in this paper, as it simplifies our analysis. 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 and Nussbaum, 2012).

2.2 Estimating OT distances with realizations

When the input distributions are discrete (or equivalently when $\mathcal{X}$ is a finite set), i.e. $\alpha = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\beta = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$, the function f_t and g_t need only to be evaluated on $(x_i)_t$ and $(y_i)_i$, which allows a proper implementation. The iterations (3) then correspond to the Sinkhorn and Knopp (1967) algorithm over the inverse scaling vectors $\boldsymbol{u}_t \triangleq (e^{-f_t(x_i)})_{i=1}^n, \boldsymbol{v}_t \triangleq (e^{-g_t(y_i)})_{i=1}^n$:

$$\boldsymbol{u}_{t+1} = \boldsymbol{K}\frac{1}{n\boldsymbol{v}_t} \quad \text{and} \quad \boldsymbol{v}_{t+1} = \boldsymbol{K}^\top \frac{1}{n\boldsymbol{u}_t} \tag{4}$$

where $\boldsymbol{K} = (e^{-C(x_i,y_i)})_{i,j=1}^n \in \mathbb{R}^{n\times n}$, and inversion is made pointwise. The Sinkhorn algorithm for OT thus operates in two phases: first, the kernel matrix $\boldsymbol{K}$ is computed, with a cost in $O(n^2 d)$, where d is the dimension of $\mathcal{X}$; then each iteration (4) costs $O(n^2)$. The online Sinkhorn algorithm that we propose mixes these two phases to accelerate convergence (see results in §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 $\mathcal{W}(\hat{\alpha}_n, \hat{\beta}_n) \to \mathcal{W}(\alpha,\beta)$. Although this is a reassuring result, the sample complexity of transport in high dimensions with low regularization remains high (see §1).

The estimation of $\mathcal{W}(\alpha,\beta)$ may be improved using several i.i.d sets of samples $(\hat{\alpha}_t)_t$ and $(\hat{\beta}_t)_t$. Those should be of reasonable size to fit in memory and may for example come from a temporal stream. Genevay, Peyré, et al., 2018 use a Monte-Carlo estimate $\hat{\mathcal{W}}(\alpha,\beta) = \frac{1}{T}\sum_{t=1}^T \mathcal{W}(\hat{\alpha}_t, \hat{\beta}_t)$. However, this yields a biased estimation as the distance $\mathcal{W}(\alpha,\beta)$ and the optimal potentials $f^\star = f^\star(\alpha,\beta)$ differ from their expectation under sampling $\mathbb{E}_{\hat{\alpha}\sim\alpha,\hat{\beta}\sim\beta}[\mathcal{W}(\hat{\alpha},\hat{\beta})]$ and $\mathbb{E}_{\hat{\alpha}\sim\alpha,\hat{\beta}\sim\beta}[f^\star(\hat{\alpha},\hat{\beta})]$. In contrast, online Sinkhorn consistently estimates the true potential functions (up to a constant) and the Sinkhorn cost.

3 OT distances from sample streams

We now introduce an online adaptation of the Sinkhorn algorithm. We construct functional estimators of $f^\star$, $g^\star$ and $\mathcal{W}(\alpha,\beta)$ using successive discrete distributions of samples $(\hat{\alpha}_t)_t$ and $(\hat{\beta}_t)_t$, where $\hat{\alpha}_t \triangleq \frac{1}{n}\sum_{i=n_t+1}^{n_{t+1}} \delta_{x_i}$, with $n_0 \triangleq 0$ and $n_{t+1} \triangleq$

$n_t + n$. The size of the mini-batch n may potentially depends on t. $(\hat{\alpha}_t)_t$ and $(\hat{\beta}_t)_t$ may be seen as mini-batches of size n within a training procedure.

3.1 Online Sinkhorn iterations

The optimization trajectory $(f_t, g_t)_t$ of the continuous Sinkhorn algorithm given by (3) is untractable as it cannot be represented in memory. The exp-potentials $u_t \triangleq \exp(-f_t)$ and $v_t \triangleq \exp(-g_t)$ are indeed infinitesimal mixtures of kernel functions $\kappa_y(\cdot) \triangleq \exp(-C(\cdot, y))$ and $\kappa_x(\cdot) \triangleq \exp(-C(x, \cdot))$.

We propose to construct finite-memory consistent estimates of u_t and v_t using principles from stochastic approximation (SA) Robbins and Monro, 1951. We cast the regularized OT problem as a root-finding problem of a function-valued operator $\mathcal{F} : \mathcal{C}_+(\mathcal{X}) \times \mathcal{C}_+(\mathcal{X}) \to \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{X})$, for which we can obtained unbiased estimates. Optimal potentials are indeed exactly the roots of

$$\mathcal{F} : (u, v) \to \Big(u(\cdot) - \int_{y \in \mathcal{X}} \frac{1}{v(y)} \kappa_y(\cdot) \mathrm{d}\beta(y), \quad v(\cdot) - \int_{x \in \mathcal{X}} \frac{1}{u(x)} \kappa_x(\cdot) \mathrm{d}\alpha(x) \Big).$$

In particular, the simultaneous Sinkhorn updates rewrites as $(u_{t+1}, v_{t+1}) = (u_t, v_t) - \mathcal{F}(u_t, v_t)$ for all t. Importantly, $\mathcal{F}$ can be evaluated without bias using two empirical measures $\hat{\alpha}$ and $\hat{\beta}$, defining

$$\hat{\mathcal{F}}_{\hat{\alpha}, \hat{\beta}}(u, v) \triangleq \Big(u(\cdot) - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{v(y_i)} \kappa_{y_i}(\cdot) \quad v(\cdot) - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{u(x_i)} \kappa_{x_i}(\cdot) \Big).$$

By construction, $\mathbb{E}_{\hat{\alpha} \sim \alpha, \hat{\beta} \sim \beta}[\hat{\mathcal{F}}_{\hat{\alpha}, \hat{\beta}}] = \mathcal{F}$, and the images of $\hat{\mathcal{F}}$ admit a representation in memory.

Randomized Sinkhorn. To make use of a stream of samples $(\hat{\alpha}_t, \hat{\beta}_t)_t$, we may simply replace $\mathcal{F}$ with $\hat{\mathcal{F}}$ in the Sinkhorn updates. This amounts to use noisy soft C-transforms in (3), as we set

$$\begin{aligned} (u_{t+1}, v_{t+1}) &\triangleq (u_t, v_t) - \hat{\mathcal{F}}_{\hat{\alpha}, \hat{\beta}}(u_t, v_t), \quad \text{i.e.} \\ \hat{f}_{t+1} &= T_{\hat{\beta}_t}(\hat{g}_t), \qquad \hat{g}_{t+1} = T_{\hat{\alpha}_t}(\hat{f}_{t+1}). \end{aligned} \tag{5}$$

$\hat{f}_t$ and $\hat{g}_t$ are defined in memory by $(y_i, \hat{g}_{t-1}(y_i))_i$ and $(x_i, \hat{f}_{t-1}(x_i))_i$. Yet the variance of the updates (5) does not decay through time, hence this *randomized Sinkhorn* algorithm does not converge. However, we show in Proposition 1 that the Markov chain $(\hat{f}_t, \hat{g}_t)_t$ converges towards a stationary distribution that is independent of the potentials $\hat{f}_0$ and $\hat{g}_0$ used for initialization.

Online Sinkhorn. To ensure convergence of $\hat{f}_t$, $\hat{g}_t$ towards some optimal pair of potentials $(f^\star, g^\star)$, one must take more cautious steps, in particular past iterates should not be discarded. We introduce a learning rate η_t in Sinkhorn

Algorithm 1 Online Sinkhorn

Input: Dist. α and β, learning weights $(\eta_t)_t$, batch sizes $(n(t))_t$ **Set** $p_i = q_i = 0$ for $i \in (0, n_1]$

for $t = 0, \dots, T-1$ **do**

Sample $(x_i)_{(n_t, n_{t+1}]} \sim \alpha$, $(y_j)_{(n_t, n_{t+1}]} \sim \beta$.

Evaluate $(\hat{f}_t(x_i))_{i=(n_t,n_{t+1}]}$, $(\hat{g}_t(y_i))_{i=(n_t,n_{t+1}]}$ using $(q_{i,t}, p_{i,t}, x_i, y_i)_{i=(0,n_t]}$ in (7).

$q_{(n_t,n_{t+1}],t+1} \leftarrow \log \frac{\eta_t}{n} + (\hat{g}_t(y_i))_{(n_t,n_{t+1}]}$, $p_{(n_t,n_{t+1}],t+1} \leftarrow \log \frac{\eta_t}{n} + (\hat{f}_t(x_i))_{(n_t,n_{t+1}]}$.

$q_{(0,n_t],t+1} \leftarrow q_{(0,n_t],t} + \log(1-\eta_t)$, $p_{(0,n_t],t+1} \leftarrow p_{(0,n_t],t} + \log(1-\eta_t)$.

Returns: $\hat{f}_T : (q_{i,T}, y_i)_{(0,n_T]}$ and $\hat{g}_T : (p_{i,T}, x_i)_{(0,n_T]}$

iterations, akin to the Robbins-Monro algorithm for finding roots of vector-valued functions:

$$(\hat{u}_{t+1}, \hat{v}_{t+1}) \triangleq (1-\eta_t)(\hat{u}_t, \hat{v}_t) - \eta_t \hat{\mathcal{F}}_{\hat{\alpha}_t, \hat{\beta}_t}(\hat{u}_t, \hat{v}_t), \quad \text{i.e.} \tag{6}$$
$$e^{-\hat{f}_{t+1}} = (1-\eta_t) e^{-\hat{f}_t} + \eta_t e^{-T_{\hat{\beta}_t}(\hat{g}_t)}$$

Each update adds new kernel functions to a non-parametric estimation of u_t and v_t. The estimates $\hat{u}_t$ and $\hat{v}_t$ are defined by weights $(p_{i,t}, q_{i,t})_{i \leqslant n_t}$ and positions $(x_i, y_i)_{i \leqslant n_t} \subseteq \mathcal{X}^2$:

$$e^{-\hat{f}_t(\cdot)} = \hat{u}_t(\cdot) \triangleq \sum_{i=1}^{n_t} \exp(q_{i,t} - C(\cdot, y_i)), \tag{7}$$
$$e^{-\hat{g}_t(\cdot)} = \hat{v}_t(\cdot) \triangleq \sum_{i=1}^{n_t} \exp(p_{i,t} - C(x_i, \cdot)).$$

The SA updates (6) yields simple vectorized updates for the weights $(p_i, q_i)_i$, leading to Algorithm 1. We perform the updates for q_i and p_i in log-space, for numerical stability reasons.

Complexity. Each iteration of online Sinkhorn has complexity $\mathcal{O}(n_t\, n)$, due to the evaluation of the distances $C(x_i, y_i)$ for all $(x_i)_{(0,n_t]}$ and $(y_i)_{(n_t,n_{t+1}]}$, and the soft C-transforms in (7). Online Sinkhorn computes a distance matrix $(C(x_i, y_j))_{i,j \leqslant n_t}$ on the fly, in parallel to updating $\hat{f}_t$ and $\hat{g}_t$. In total, its computation cost after drawing n_t samples is $\mathcal{O}(n_t^2)$. Its memory cost is $\mathcal{O}(n_t)$; it increases with iterations, which is a requirement for consistent estimation. Randomized Sinkhorn with constant batch-sizes n has a memory cost of $\mathcal{O}(n)$ and a single-iteration computational cost of $\mathcal{O}(n^2)$.

3.2 Refinements

Estimating Sinkhorn distance. As we will see in §4, the iterations (6) only estimate potential functions up to a constant. This is sufficient for minimizing a

loss function involving a Sinkhorn distance (e.g. for model training or barycenter estimation (Staib et al., 2017)), as backpropagating through the Sinkhorn distance relies only on the gradients of the potentials $\nabla_x f^\star(\cdot)$, $\nabla_y g^\star(\cdot)$ (e.g. Cuturi and Peyré, 2018). With extra $\mathcal{O}(n_t^2)$ operations, $(\hat{f}_t, \hat{g}_t)$ may be used to estimate $\mathcal{W}(\alpha, \beta)$ through a final soft C-transform:

$$\hat{\mathcal{W}}_t \triangleq \frac{1}{2}\Big(\langle \bar{\alpha}_t,\, f_t + T_{\bar{\alpha}_t}(\hat{g}_t)\rangle + \langle \bar{\beta}_t,\, \hat{g}_t + T_{\bar{\alpha}_t}(f_t)\rangle\Big),$$

where $\bar{\alpha}_t \triangleq \frac{1}{n_t}\sum_{i=1}^{n_t} \delta_{x_i}$ and $\bar{\beta}_t$ are formed of all previously observed samples.

Fully-corrective scheme. The potentials $\hat{f}_t$ and $\hat{g}_t$ may be improved by refitting the weights $(p_i)_{(0,n_t]}$, $(q_j)_{(0,n_t]}$ based on all previously seen samples. For this, we update $\hat{f}_{t+1} = T_{\bar{\beta}_t}(g_t)$ and $\hat{g}_{t+1} = T_{\bar{\alpha}_t}(f_t)$. This reweighted scheme (akin to the fully-corrective Frank-Wolfe scheme from Lacoste-Julien and Jaggi, 2015) has a cost of $\mathcal{O}(n_t^2)$ per iteration. It requires to keep in memory (or recompute on-the-fly) the whole distance matrix. Fully-corrective online Sinhorn enjoys similar convergence properties as regular online Sinkhorn, and permits the use of non-increasing batch-sizes—see §B.1. In practice, it can be used every k iterations, with k increasing with t. Combining partial and full updates can accelerate the estimation of Sinkhorn distances (see §5.2).

Finite samples. Finally, we note that our algorithm can handle 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 update at each iterations a set of coordinates of size $n < N$. The resulting algorithm is a *subsampled* Sinkhorn algorithm for histograms, which is detailed in §B.2, Algorithm 3. We show in §5 that it is useful to accelerate the first phase of the Sinkhorn algorithm.

4 Convergence analysis

We show a stationary distribution convergence property for the randomized Sinkhorn algorithm, an approximate convergence property for the online Sinkhorn algorithm with fixed batch-size and an exact convergence result for online Sinkhorn with increasing batch sizes, with asymptotic convergence rates. We make the following classical assumption on the cost regularity and compactness of α and β.

Assumption 1. *The cost $C : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is L-Lipschitz, and $\mathcal{X}$ is compact.*

4.1 Randomized Sinkhorn

We first state a result concerning the randomized Sinkhorn algorithm (5), proved in §A.2.

Proposition 1. *Under Assumption 1, the randomized Sinkhorn algorithm (5) yields a time-homogeneous Markov chain* $(\hat{f}_t, \hat{g}_t)_t$ *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 and Freedman (1999) convergence theorem on iterated random functions which are contracting on average. We use the fact that $T_{\hat{\beta}}(\cdot)$ and $T_{\hat{\alpha}}(\cdot)$ are *uniformly* contracting, independently of the distributions $\hat{\alpha}$ and $\hat{\beta}$, for the variational norm $\|\cdot\|_{\text{var}}$. Using the law of large number for Markov chains (Breiman, 1960), the (tractable) average $\frac{1}{t}\sum_{s=1}^{t} \exp(-\bar{f}_s)$ converges almost surely to $\mathbb{E}[e^{-f_\infty}] \in \mathcal{C}(\mathcal{X})$. This expectation verifies the functional equations

$$\mathbb{E}[e^{-f_\infty}] = \int_y \mathbb{E}[e^{g_\infty(y) - C(\cdot, y)}] \mathrm{d}\beta(y) \quad \mathbb{E}[e^{-g_\infty}] = \int_x \mathbb{E}[e^{f_\infty(x) - C(x, \cdot)}] \mathrm{d}\alpha(x)$$

These equations are close to the Sinkhorn fixed point equations, and get closer as ε increases, since $\varepsilon \mathbb{E}[\exp(\pm f_\infty / \varepsilon)] \to \mathbb{E}[\pm f_\infty]$ as $\varepsilon \to \infty$. Running the random Sinkhorn algorithm with averaging fails to provide exactly the dual solution, but solves an approximate problem.

4.2 Online Sinkhorn

We make the following Robbins and Monro (1951) assumption on the weight sequence. We then state an approximate convergence result for the online Sinkhorn algorithm with fixed batch-size $n(t) = n$.

Assumption 2. $(\eta_t)_t$ *is such that* $\sum \eta_t = \infty$ *and* $\sum \eta_t^2 < \infty$*,* $0 \leqslant \eta_t \leqslant 1$ *for all* $t > 0$*.*

Proposition 2. *Under Assumption 1 and 2, the online Sinkhorn algorithm (Algorithm 1) yields a sequence* (f_t, g_t) *that reaches a ball centered around* $f^\star, g^\star$ *for the variational norm* $\|\cdot\|_{\text{var}}$*. Namely, there exists* $T > 0$*,* $A > 0$ *such that for all* $t > T$*, almost surely*

$$\|f_t - f^\star\|_{\text{var}} + \|g_t - g^\star\|_{\text{var}} \leqslant \frac{A}{\sqrt{n}}.$$

The proof is reported in §A.3. It is not possible to ensure the convergence of online Sinkhorn with constant batch-size. This is a fundamental difference with other SA algorithms, e.g. SGD on strongly convex objectives (see Moulines and Bach, 2011). This stems from the fact that the metric for which $\mathrm{Id} - \mathcal{F}$ is contracting is not a Hilbert norm. The constant A depends on L, the diameter of $\mathcal{X}$ and the regularity of potentials $f^\star$ and $g^\star$, but not on the dimension. It behaves like $\exp(\frac{1}{\varepsilon})$ when $\varepsilon \to 0$. Fortunately, we can show the almost sure convergence of the online Sinkhorn algorithm with slightly increasing batch-size $n(t)$ (that may grow arbitrarily slowly for $\eta_t = \frac{1}{t}$), as specified in the following assumption.

Assumption 3. *For all* $t > 0$, $n(t) = \frac{B}{w_t^2} \in \mathbb{N}$ *and* $0 \leqslant \eta_t \leqslant 1$. $\sum w_t \eta_t < \infty$ *and* $\sum \eta_t = \infty$.

Proposition 3. *Under Assumption 1 and 3, the online Sinkhorn algorithm converges almost surely:*

$$\|\hat{f}_t - f^\star\|_{\text{var}} + \|\hat{g}_t - g^\star\|_{\text{var}} \to 0.$$

The proof is reported in §A.4. It relies on a uniform law of large number for functions (Van der Vaart, 2000, chapter 19) and on the uniform contractivity of soft C-transform operator (e.g. Vialard, 2019, Proposition 19). Consistency of the iterates is an original property—Genevay, Cuturi, et al., 2016 only show convergence of the OT value. Finally, using bounds from Moulines and Bach, 2011, we derive asymptotic rates of convergence for online Sinkhorn (see §A.5), with respect to the number of observed samples N. We write $\delta_N = \|\hat{f}_{t(N)} - f^\star\|_{\text{var}} + \|\hat{g}_{t(N)} - g^\star\|_{\text{var}}$, where $t(N)$ is the iteration number for which $n_t > N$ samples have been observed.

Proposition 4. *For all* $\iota \in (0, 1)$, $S > 0$ *and* $B \in \mathbb{N}^\star$, *setting* $\eta_t = \frac{S}{t^{1-\iota}}$, $n(t) = \lceil B t^{4\iota} \rceil$, *there exists* $D > 0$ *independant of* N *and* $N_0 > 0$ *such that, for all* $N > N_0$, $\delta_N \leqslant \frac{D}{N^{\frac{1-\iota}{1+4\iota}}}$.

Online Sinkhorn thus provides estimators of potentials whose asymptotic sample complexity in variational norm is arbitrarily close to $\mathcal{O}(\frac{1}{N})$. To the best of our knowledge, this is an original property. It also results in a distance estimator $\hat{\mathcal{W}}_N$ whose complexity is arbitrarily close to $\mathcal{O}(\frac{1}{\sqrt{N}})$, recovering existing asymptotic rates from Genevay, Chizat, et al., 2019, for any Lipschitz cost. We derive non-asymptotic rates in §A.5 (see (19)), which make explicit the bias-variance trade-off when choosing the step-sizes and batch-sizes. We also give the explicit form of D; it does not depend on the dimension. For low ε, D is proportional to $\exp(\frac{2}{\varepsilon})$; the bound is therefore vacuous for $\varepsilon \to 0$. Note that using growing batch-sizes amounts to increase the budget of a single iteration over time: the overall computational complexity after seeing N samples is always $\mathcal{O}(N^2)$.

Batch-sizes and step-sizes. To provide practical guidance on choosing rates in batch-sizes $n(t)$ and step-sizes η_t, we can parametrize $\eta_t = \frac{1}{t^a}$ and $n(t) = Bt^b$ and study what is implied by Assumption 3 and Assumption 4. We summarize the schedules for which convergence is guarantees in Table 1. Note that in practice, it is useful to replace t by $(1 + r\,t)$ in these schedules. We set $r = 0.1$ in all experiments.

Mirror descent interpretation. Online Sinkhorn can be interpreted as a non-convex stochastic mirror-descent, as detailed in Appendix D. It provides an original interpretation of the Sinkhorn algorithm, different from recent work (Léger, 2019; Mishchenko, 2019).

Table 1: Schedules of batch-sizes and learning rates that ensures online Sinkhorn convergence.

Param. schedule	Online Sinkhorn	Fully-corrective online Sinkhorn
Batch size $n(t) = Bt^b$	$0 < b$	$0 \leqslant b$
Step size $\eta_t = \frac{1}{t^a}$	$a \geqslant 1 - \frac{b}{2}$	$\begin{cases} a > \frac{1}{2} - \frac{b}{2} & \text{and} \quad b < 1 \\ a \geqslant 0 & \text{and} \quad b \geqslant 1 \end{cases}$

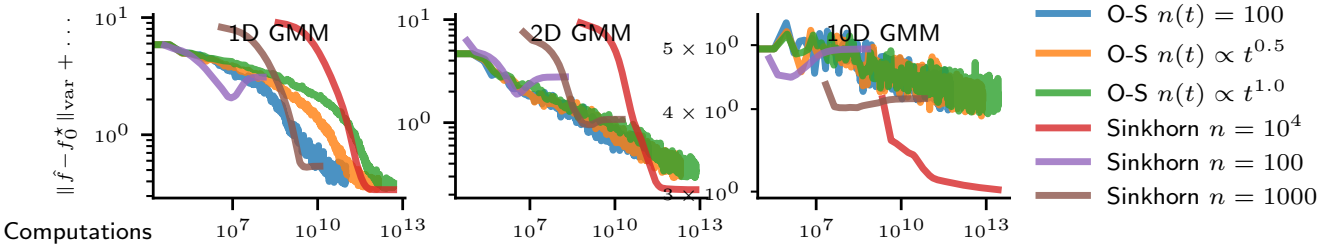


Figure 1: Online Sinkhorn consistently estimate the true regularized OT potentials. Convergence here is measured in term of distance with potentials evaluated on a "test" grid of size $n = 10^4$. Online-Sinkhorn can estimate potentials faster than sampling then scaling the cost matrix.

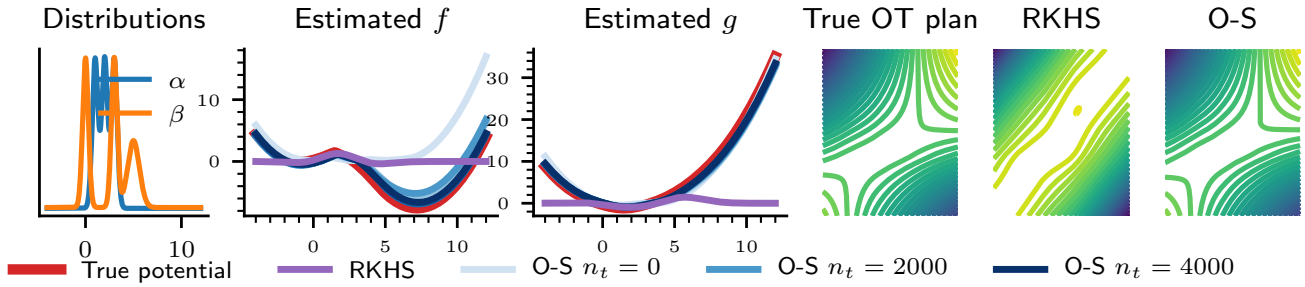


Figure 2: Online Sinhkorn finds the correct potentials over all space, unlike SGD over a RKHS parametrization of the potentials. The plan is therefore correctly estimated everywhere.

5 Numerical experiments

The major purpose of online Sinkhorn (OS) is to handle OT between continuous distributions. We first show that it is a valid alternative to applying Sinkhorn on a single realization of continuous distributions, using examples of Gaussian mixtures of varying dimensions. We then illustrate that OS is able to estimate

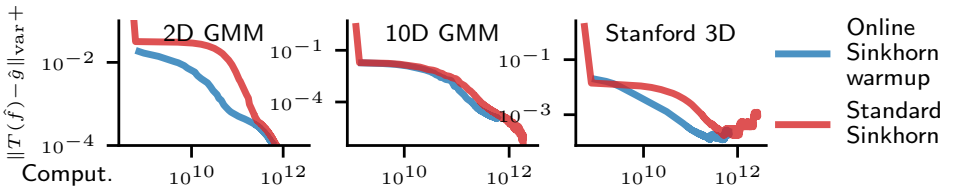
precisely Kantorovich dual potentials, significantly improving the result obtained using SGD with RKHS expansions (Genevay, Cuturi, et al., 2016). Finally, we show that OS is an efficient warmup strategy to accelerate Sinkhorn for discrete problems on several real and synthetic datasets.

5.1 Continuous potential estimation with online Sinkhorn

Data and quantitative evaluation. We measure the performance of our algorithm in a continuous setting, where α and β are parametric distributions (Gaussian mixtures in 1D, 2D and 10D, with 3, 3 and 5 modes, so that $C_{\max} \sim 1$), from which we draw samples. In the absence of reference potentials $(f^\star, g^\star)$ (which cannot be computed in closed form), we compute "test" potentials $(f_0^\star, g_0^\star)$ on realizations $\hat{\alpha}_0$ and $\hat{\beta}_0$ of size 10000, using Sinkhorn. We then compare OS to Sinkhorn runs of various size , trained on realizations $N = (100, 1000, 10000)$ independent of the reference grid (to avoid reducing the problem to a discrete problem between $\hat{\alpha}_0$ and $\hat{\beta}_0$). To measure convergence, we compute $\delta_t = \|\hat{f}_t - f_0^\star\|_{\text{var}} + \|\hat{g}_t - g_0^\star\|_{\text{var}}$, evaluated on the grid defined by $\hat{\alpha}_0$ and $\hat{\beta}_0$, which constitutes a Monte-Carlo approximation of the error. We evaluate OS with and without full-correction, with different batch-size schedules (see §C.1), as well as the randomized Sinkhorn algorithm. Quantitative results are average over 5 runs. We report quantitative results for $\varepsilon = 10^{-2}$ and non fully-corrective online Sinkhorn in the main text, and all other curves in Supp. Fig. 4. In Supp. Fig. 7, we also report results for OT between Gaussians, which is a simpler and less realistic setup, but for which closed-form expressions of the potentials are known Janati et al., 2020.

Comparison to SGD. For qualitative illustration, on the 1D and 2D problem, we consider the main existing competing approach (Genevay, Cuturi, et al., 2016), in which $f_t(\cdot)$ is parametrized as $\sum_{i=1}^{n_t} \alpha_t \kappa(\cdot, x_i)$ (and similarly for g_t), where κ is a reproducing kernel (typically a Gaussian). This differs significantly from online Sinkhorn, where we express e^{-f_t} as a Gaussian mixture. The dual problem (3) is solved using SGD, with convergence guarantees on the dual energy. As advocated by the authors, we run a grid search over the bandwidth parameter σ of the Gaussian kernel to select the best performing runs.

Earlier potential convergence. We study convergence curves in Fig. 1, comparing algorithms at equal number of multiplications. OS outperforms or matches Sinkhorn for $N = 100$ and $N = 1000$ on the three problems; it approximately matches the performance of Sinkhorn on $N = 10000$ new iterates on the 1D and 2D problems. On the two low-dimensional problems, online Sinkhorn converges faster than Sinkhorn at the beginning. Indeed, it initiates the computation of the potentials early, while the Sinkhorn algorithm must wait for the cost matrix to be filled. This leads us to study online Sinkhorn as a catalyser of Sinkhorn in the next paragraph. OS convergence is slower (but is still noticeable) for the higher dimensional problem. Fully-corrective OS performs



$-\log \varepsilon$	4	3	2	1
Stanford	5.3x	4.4x	2.3x	1.2x
10D GMM	1.4x	1.5x	1.3x	1.2x
2D GMM	17x	3.7x	1.3x	2.0x

Figure 3: Online Sinkhorn allows to warmup Sinkhorn during the evaluation of the cost matrix, and to speed discrete optimal transport. Table 2: Speed-ups provided by OS vs S to reach a 10^{-3} precision.

better in this case (see Supp. Fig. 5). We also note that randomized Sinkhorn with batch-size N performs on par with Sinkhorn of size N (Supp. Fig. 6).

Better-extrapolated potentials. As illustrated in Fig. 2, in 1D, online Sinkhorn refines the potentials $(\hat{f}_t, \hat{g}_t)_t$ until convergence toward $(f^\star, g^\star)$. Supp. Fig. 8 shows a visualisation for 2D GMM. As the parametrization (7) is adapted to the dual problem, the algorithm quickly identifies the correct shape of the optimal potentials—as predicted by Proposition 3. In particular, OS estimates potentials with much less errors than SGD in a RKHS in areas where the mass of α and β is low. This allows to consistently estimate the transport plan, which cannot be achieved using SGD. SGD did not converge for $\varepsilon < 10^{-1}$, while online Sinkhorn remains stable. OS does not require to set a bandwidth.

5.2 Accelerating Sinkhorn with online Sinkhorn warmup

The discrete Sinkhorn algorithm requires to compute the full cost matrix $\boldsymbol{C} \triangleq (C(x_i, y_i))_{i,j}$ of size $N \times N$, prior to estimating the potentials $\boldsymbol{f}_1 \in \mathbb{R}^N$ and $\boldsymbol{g}_1 \in \mathbb{R}^N$ by a first C-transform. In contrast, online Sinkhorn can progressively compute this matrix while computing first sketches of the potentials. The extra cost of estimating the initial potentials without full-correction is simply $2N^2$, i.e. similar to filling-up $\boldsymbol{C}$. We therefore assess the performance of *online Sinkhorn as Sinkhorn warmup* 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 $\boldsymbol{C}$ is completely evaluated. From then, the subsequent potentials are obtained using full Sinkhorn updates. We consider the GMMs of §5.1, as well as a 3D dragon from Stanford 3D scans Turk and Levoy, 1994 and a sphere of size $N = 12000$. We measure convergence using the error $\|T_\alpha(\hat{f}_t) - \hat{g}_t\|_{\text{var}} + \|T_\beta(\hat{g}_t) - \hat{f}_t\|_{\text{var}}$, evaluated on the support of α and β; this error goes to 0. We use $n(t) = \frac{N}{100}(1 + 0.1t)^{1/2}$—results vary little with the exponent.

Results. We report convergence curves for $\varepsilon = 10^{-3}$ in Fig. 3, and speed-ups due to OS in Table 1. Convergence curves for different ε are reported in Supp. Fig. 9. The proposed scheme provides an improvement upon the

standard Sinkhorn algorithm. After $N^2 d$ computations (the cost of estimating the full matrix C), both the function value and distance to optimum are lower using OS: the full Sinkhorn updates then relay the online updates, using an accurate initialization of the potentials. The *OS warmed-up* Sinkhorn algorithm then maintains its advantage over the standard Sinkhorn algorithm during the remaining iterations. The speed gain increases as ε reduces and the OT problem becomes more challenging. Sampling without replacement brings an additional speed-up.

6 Conclusion

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. We have drawn links between regularized OT and stochastic approximation. This opens promising avenues to study convergence rates of continuous variants of Sinkhorn's iterations. Future work will refine the complexity constants and design adaptive non-parametric potential estimations.

7 Acknowledgements

This work was supported by the European Research Council (ERC project NORIA). A.M thanks Anna Korba for helpful discussions on mirror descent algorithms, and Thibault Séjourné for proof-reading and relevant references.

References

Alber, Ya. I., C. E. Chidume, and Jinlu Li (2012). "Stochastic Approximation Method for Fixed Point Problems". In: *Applied Mathematics* 03.12, pp. 2123–2132.

Altschuler, Jason, Jonathan Niles-Weed, and Philippe Rigollet (2017). "Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration". In: *Advances in Neural Information Processing Systems*.

Arjovsky, Martin, Soumith Chintala, and Léon Bottou (2017). "Wasserstein Generative Adversarial Network". In: *Proceedings of the International Conference on Machine Learning*.

Beck, Amir and Marc Teboulle (2003). "Mirror descent and nonlinear projected subgradient methods for convex optimization". In: *Operations Research Letters* 31.3, pp. 167–175.

Benamou, Jean-David, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré (2015). "Iterative Bregman projections for regularized transportation problems". In: *SIAM Journal on Scientific Computing* 37.2, A1111–A1138.

Breiman, Leo (1960). "The Strong Law of Large Numbers for a Class of Markov Chains". In: *The Annals of Mathematical Statistics* 31.3, pp. 801–803.

Canas, Guillermo and Lorenzo Rosasco (2012). "Learning probability measures with respect to optimal transport metrics". In: *Advances in Neural Information Processing Systems*.

Celeux, Gilles and Jean Diebolt (1992). "A stochastic approximation type EM algorithm for the mixture problem". In: *Stochastics and Stochastic Reports* 41.1, pp. 119–134.

Chizat, Lenaic (2019). "Sparse Optimization on Measures with Over-parameterized Gradient Descent". In: *arXiv preprint arXiv:1907.10300v1*.

Courty, Nicolas, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy (2016). "Optimal transport for domain adaptation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.9, pp. 1853–1865.

Cuturi, Marco (2013). "Sinkhorn Distances: Lightspeed Computation of Optimal Transport". In: *Advances in Neural Information Processing Systems*.

Cuturi, Marco and Gabriel Peyré (2018). "Semidual Regularized Optimal Transport". In: *SIAM Review* 60.4, pp. 941–965.

Dalalyan, Arnak S and Avetik Karagulyan (2019). "User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient". In: *Stochastic Processes and their Applications* 129.12, pp. 5278–5311.

Diaconis, Persi and David Freedman (1999). "Iterated random functions". In: *SIAM Review* 41.1, pp. 45–76.

Dudley, R. M. (1969). "The Speed of Mean Glivenko-Cantelli Convergence". In: *The Annals of Mathematical Statistics* 40.1, pp. 40–50.

Fatras, Kilian, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty (2019). "Learning with minibatch Wasserstein: asymptotic and gradient properties". In: *arXiv preprint arXiv:1910.04091*.

Feydy, Jean, Thibault Séjourné, François-Xavier Vialard, S. Amari, Alain Trouvé, and Gabriel Peyré (2019). "Interpolating between Optimal Transport and MMD using Sinkhorn Divergences". In: *International Conference on Artificial Intelligence and Statistics*.

Frogner, Charlie, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio (2015). "Learning with a Wasserstein loss". In: *Advances in Neural Information Processing Systems*.

Genevay, Aude, Lenaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré (2019). "Sample Complexity of Sinkhorn divergences". In: *International Conference on Artificial Intelligence and Statistics*.

Genevay, Aude, Marco Cuturi, Gabriel Peyré, and Francis Bach (2016). "Stochastic Optimization for Large-scale Optimal Transport". In: *Advances in Neural Information Processing Systens*, pp. 3432–3440.

Genevay, Aude, Gabriel Peyré, and Marco Cuturi (2018). "Learning Generative Models with Sinkhorn Divergences". In: *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617.

Goldberg, Andrew V and Robert E Tarjan (1989). "Finding minimum-cost circulations by canceling negative cycles". In: *Journal of the ACM* 36.4, pp. 873–886.

Hsieh, Ya-Ping, Chen Liu, and Volkan Cevher (2018). "Finding Mixed Nash Equilibria of Generative Adversarial Networks". In: *arXiv preprint arXiv:1811.02002*.

Janati, Hicham, Boris Muzellec, Gabriel Peyré, and Marco Cuturi (2020). "Entropic Optimal Transport between (Unbalanced) Gaussian Measures has a Closed Form". In: *arXiv:2006.02572 [math, stat]*. arXiv: 2006.02572.

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

Lacoste-Julien, Simon and Martin Jaggi (2015). "On the global linear convergence of Frank-Wolfe optimization variants". In: *Advances in Neural Information Processing Systems*.

Léger, Flavien (2019). "A gradient descent perspective on Sinkhorn". In: *arXiv preprint arXiv:2002.03758*.

Lemmens, Bas and Roger Nussbaum (2012). *Nonlinear Perron–Frobenius Theory*. Cambridge: Cambridge University Press.

Mairal, Julien (2013). "Stochastic majorization-minimization algorithms for large-scale optimization". In: *Advances in Neural Information Processing Systems*.

Mensch, Arthur, Mathieu Blondel, and Gabriel Peyré (2019). "Geometric Losses for Distributional Learning". In: *Proceedings of the International Conference on Machine Learning*.

Mérigot, Quentin (2011). "A multiscale approach to optimal transport". In: *Computer Graphics Forum*. Vol. 30. 5. Wiley Online Library, pp. 1583–1592.

Mishchenko, Konstantin (2019). "Sinkhorn Algorithm as a Special Case of Stochastic Mirror Descent". In: *arXiv preprint arXiv:1909.06918*.

Moulines, Eric and Francis Bach (2011). "Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning". In: *Advances in Neural Information Processing Systems*, pp. 451–459.

Peyré, Gabriel and Marco Cuturi (2019). "Computational optimal transport". In: *Foundations and Trends® in Machine Learning* 11.5-6, pp. 355–607.

Robbins, Herbert and Sutton Monro (1951). "A stochastic approximation method". In: *The Annals of Mathematical Statistics*, pp. 400–407.

Santambrogio, Filippo (2015). "Optimal transport for applied mathematicians". In: *Birkäuser, NY* 55.58-63, p. 94.

Seguy, Vivien, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel (2018). "Large-scale optimal transport and mapping estimation". In: *International Conference on Learning Representations*.

Sinkhorn, Richard (1964). "A relationship between arbitrary positive matrices and doubly stochastic matrices". In: *The Annals of Mathematical Statistics* 35, pp. 876–879.

Sinkhorn, Richard and Paul Knopp (1967). "Concerning nonnegative matrices and doubly stochastic matrices". In: *Pacific Journal of Mathematics* 21.2, pp. 343–348.

Staib, Matthew, Sebastian Claici, Justin M Solomon, and Stefanie Jegelka (2017). "Parallel streaming Wasserstein barycenters". In: *Advances in Neural Information Processing Systems*.

Turk, Greg and Marc Levoy (1994). "Zippered polygon meshes from range images". In: *Proceedings of the 21st annual conference on Computer graphics and interactive techniques*, pp. 311–318.

Van der Vaart, Aad W. (2000). *Asymptotic statistics*. Cambridge University Press.

Vialard, François-Xavier (2019). "An elementary introduction to entropic regularization and proximal methods for numerical optimal transport". In:

Weed, Jonathan and Francis Bach (2019). "Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance". In: *Bernoulli* 25.4A, pp. 2620–2648.

A Proofs

We first introduce two useful known lemmas, and prove the propositions in their order of appearance.

A.1 Useful lemmas

First, under Assumption 1, we note that the soft C-transforms are uniformly contracting on the distribution space $\mathcal{P}(\mathcal{X})$. This is clarified in the following lemma, extracted from Vialard (2019), Proposition 19. We refer the reader to the original references for proofs.

Lemma 1. *Unser Assumption 1, let* $\kappa = 1-\exp(-L\mathrm{diam}(\mathcal{X}))$. *For all* $\hat{\alpha} \in \mathcal{P}(\mathcal{X})$ *and* $\hat{\beta} \in \mathcal{P}(\mathcal{X})$, *for all* $f, f', g, g' \in \mathcal{C}(\mathcal{X})$,

$$\|T_{\hat{\alpha}}(f') - T_{\hat{\alpha}}(f')\|_{\mathrm{var}} \leqslant \kappa \|f - f'\|_{\mathrm{var}}, \quad \|T_{\hat{\beta}}(g) - T_{\hat{\beta}}(g')\|_{\mathrm{var}} \leqslant \kappa \|g - g'\|_{\mathrm{var}}.$$

We will also need a uniform law of large numbers for functions. The following lemma is a consequence of Example 19.7 and Lemma 19.36 of Van der Vaart (2000), and is copied in Lemma B.6 in Mairal (2013).

Lemma 2. *Under Assumption 1, let* $(f_t)_t$ *be an i.i.d sequence in* $\mathcal{C}(\mathcal{X})$, *such that* $\mathbb{E}[f_0] = f \in \mathcal{C}(\mathcal{X})$. *Then there exists* $A > 0$ *such that, for all* $n > 0$,

$$\mathbb{E} \sup_{x \in \mathcal{X}} |\frac{1}{n} \sum_{i=1}^{n} f_i(x) - f(x)| \leqslant \frac{A}{\sqrt{n}}.$$

Finally, we need a result on running averages using the sequence $(\eta_t)_t$. The following result stems from a simple Abel transform of the law of large number, and is established by Mairal (2013), Lemma B.7.

Lemma 3. *Let* $(\eta_t)_t$ *be a sequence of weights meeting Assumption 2. Let* $(X_t)_t$ *be an i.i.d sequence of real-valued random variables with existing first moment* $\mathbb{E}[X_0]$. *We consider the sequence* $(\bar{X}_t)_t$ *defined by* $\bar{X}_0 \triangleq X_0$ *and*

$$\bar{X}_t \triangleq (1 - \eta_t)\bar{X}_{t-1} + \eta_t X_t.$$

Then $\bar{X}_t \to_{t \to \infty} \mathbb{E}[X_0]$.

A.2 Proof of Proposition 1

Proof. We use Theorem 1 from Diaconis and Freedman (1999). For this, we simply note that the space $\mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{X})$ in which the chain $x_t \triangleq (f_t, g_t)_t$, endowed with the metric $\rho((f_1, g_1), (f_2, g_2)) = \|f_1 - f_2\|_{\mathrm{var}} + \|g_1 - g_2\|_{\mathrm{var}}$, is complete and separable (the countable set of polynomial functions are dense in this space, for example). We consider the operator $A_\theta \triangleq T_{\hat{\beta}}(T_{\hat{\alpha}}(\cdot))$. $\theta \triangleq (\hat{\alpha}, \hat{\beta})$ denotes the random variable that is sampled at each iteration. We have the following recursion:

$$x_{t+2} = A_{\theta_t}(x_t).$$

From Lemma 1, for all $\hat{\alpha} \in \mathcal{P}(\mathcal{X})$, $\hat{\beta} \in \mathcal{P}(\mathcal{X})$, A_θ with $\theta = (\hat{\alpha}, \hat{\beta})$ is contracting, with module $\kappa_\theta < \kappa < 1$. Therefore

$$\int_\theta \kappa_\theta \mathrm{d}\mu(\theta) < 1, \qquad \int_\theta \log \kappa_\theta \mathrm{d}\mu(\theta) < 0.$$

Finally, we note, for all $f \in \mathcal{C}(\mathcal{X})$

$$\|T_\beta(T_{\hat{\alpha}}(f))\|_\infty \leqslant \|f\|_\infty + 2 \max_{x,y \in \mathcal{X}} C(x,y),$$

therefore $\rho(A_\theta(x_0), x_0) \leqslant 2\|x_0\|_\infty + 2\max_{x,y\in\mathcal{X}} C(x,y)$ for all θ $(\hat{\alpha}, \hat{\beta})$. The regularity condition of the theorem are therefore met. Each of the induced Markov chains $(f_{2t}, g_{2t})_t$ and $(f_{2t+1}, g_{2t+1})_t$ has a unique stationary distribution. These stationary distributions are the same: the stationary distribution is independent of the initialisation and both sequences differs only by their initialisation. Therefore $(f_t, g_t)_t$ have a unique stationary distribution (F_∞, G_∞). □

A.3 Proof of Proposition 2

For presentation purpose, we first show that the "slowed-down" online Sinkhorn algorithm converges in the absence of noise. We then turn to prove Proposition 2.

A.3.1 Noise-free online Sinkhorn

Proposition 5. *We suppose that $\hat{\alpha}_t = \alpha$, $\hat{\beta}_t = \beta$ for all t. Then the updates (6) yields a (deterministic) sequence $(f_t, g_t)_t$ such that*

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

Note that, as 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.

The "slowed-down" Sinkhorn iterations converge toward an optimal potential couple, up to a constant factor: this stems from the fact that we apply contractions in the space $(\mathcal{C}(\mathcal{X}), \|\cdot\|_{\mathrm{var}})$ with a contraction factor that decreases sufficiently slowly.

Proof. We write $(f_t, g_t)_t$ the sequence of iterates. Given a pair of optimal potentials $(f^\star, g^\star)$, we write $u_t \triangleq f_t - f^\star$, $v_t \triangleq g_t - g^\star$, $u_t^T \triangleq T_\alpha(f_t) - g^\star$ and $v_t^T \triangleq T_\alpha(g_t) - f^\star$. For all $t > 0$, we observe that

$$\begin{aligned}
\max u_{t+1} &= -\log \min \exp(-u_{t+1}) \\
&= -\log\left(\min\left((1-\eta_t)\exp(-u_t) + \eta_t \exp(-v_t^T)\right)\right) \\
&\leqslant -\log\left((1-\eta_t)\min\exp(-u_t) + \eta_t \min\exp(-v_t^T)\right) \\
&\leqslant -(1-\eta_t)\log\min\exp(-u_t) - \eta_t \log\min\exp(-v_t^T) \\
&= (1-\eta_t)\max u_t + \eta_t \max v_t^T,
\end{aligned}$$

where we have used the algorithm recursion on the second line, $\min f + g \geqslant \min f + \min g$ on the third line and Jensen inequality on the fourth line. Similarly

$$\min u_{t+1} \geqslant (1-\eta_t) \min u_t + \eta_t \min v_t^T,$$

and mirror inequalities hold for v_t. Summing the four inequalities, we obtain

$$\begin{aligned} e_{t+1} &\triangleq \|u_{t+1}\|_{\mathrm{var}} + \|v_{t+1}\|_{\mathrm{var}} \\ &= \max u_{t+1} - \min u_{t+1} + \max v_{t+1} - \min v_{t+1} \\ &\leqslant (1-\eta_t)(\|u_t\|_{\mathrm{var}} + \|v_t\|_{\mathrm{var}}) + \eta_t(\|u_t^T\|_{\mathrm{var}} + \|v_t^T\|_{\mathrm{var}}), \\ &\leqslant (1-\eta_t)(\|u_t\|_{\mathrm{var}} + \|v_t\|_{\mathrm{var}}) + \eta_t\kappa(\|u_t\|_{\mathrm{var}} + \|v_t\|_{\mathrm{var}}), \end{aligned}$$

where we use the contractivity of the soft-C-transform, that guarantees that there exists $\kappa < 1$ such that $\|v_t^T\|_{\mathrm{var}} \leqslant \kappa\|v_t\|_{\mathrm{var}}$ and $\|u_t^T\|_{\mathrm{var}} \leqslant \kappa\|u_t\|_{\mathrm{var}}$ (Peyré and Cuturi, 2019).

Unrolling the recursion above, we obtain

$$\log e_t = \sum_{s=1}^{t} \log(1 - \eta_t(1-\kappa)) + \log(e_0) \to -\infty,$$

provided that $\sum \eta_t = \infty$. The proposition follows. □

Proof of Proposition 2. For discrete realizations $\hat{\alpha}$ and $\hat{\beta}$, we define the perturbation terms

$$\varepsilon_{\hat{\beta}}(\cdot) \triangleq f^\star - T_{\hat{\beta}}(g^\star), \qquad \iota_{\hat{\alpha}}(\cdot) \triangleq g^\star - T_{\hat{\alpha}}(f^\star),$$

so that the updates can be rewritten as

$$\exp(-f_{t+1} + f^\star) = (1-\eta_t)\exp(-f_t + f^\star) + \eta_t \exp(-T_{\hat{\beta}_t}(g_t) + T_{\hat{\beta}_t}(g^\star) + \varepsilon_{\hat{\beta}_t})$$

$$\exp(-g_{t+1} + g^\star) = (1-\eta_t)\exp(-g_t + g^\star) + \eta_t \exp(-T_{\hat{\alpha}_t}(f_t) + T_{\hat{\alpha}_t}(f^\star) + \iota_{\hat{\alpha}_t}).$$

We denote $u_t \triangleq -f_t + f^\star$, $v_t \triangleq -g_t + g^\star$, $u_t^T \triangleq T_{\hat{\beta}_t}(f_t) - T_{\hat{\beta}_t}(f^\star)$, $v_t^T \triangleq T_{\hat{\beta}_t}(g_t) - T_{\hat{\beta}_t}(g^\star)$. Reusing the same derivations as in the proof of Proposition 5, we obtain

$$\begin{aligned} \|u_{t+1}\|_{\mathrm{var}} &\leqslant (1-\eta_t)\|u_t\|_{\mathrm{var}} \\ &\quad + \eta_t \log\big(\max_{x,y\in\mathcal{X}} \exp(\varepsilon_{\hat{\beta}_t}(x) - \varepsilon_{\hat{\beta}_t}(y))\exp(v_t^T(x) - v_t^T(y))\big) \\ &\leqslant (1-\eta_t)\|u_t\|_{\mathrm{var}} + \eta_t\|v_t^T\|_{\mathrm{var}} + \eta_t\|\varepsilon_{\hat{\beta}_t}\|_{\mathrm{var}}, \end{aligned}$$

where we have used $\max_x f(x)g(x) \leqslant \max_x f(x) \max_x f(x)$ on the second line. Therefore, using the contractivity of the soft C-transform,

$$e_{t+1} \leqslant (1-\tilde{\eta}_t)e_t + \frac{\tilde{\eta}_t}{1-\kappa}(\|\varepsilon_{\hat{\beta}_t}\|_{\mathrm{var}} + \|\iota_{\hat{\alpha}_t}\|_{\mathrm{var}}), \tag{8}$$

where we set $e_t \triangleq \|u_t\|_{\text{var}} + \|v_t\|_{\text{var}}$, $\tilde{\eta}_t = \eta_t(1-\kappa)$ and κ is set to be the biggest contraction factor over all empirical realizations $\hat{\alpha}_t$, $\hat{\beta}_t$ of the distributions α and β. It is upper bounded by $1 - e^{-L\text{diam}(\mathcal{X})}$, thanks to Assumption 1 and Lemma 1.

The realizations $\hat{\beta}_t$ and $\hat{\alpha}_t$ are sampled according to the same distribution $\hat{\alpha}$ and $\hat{\beta}$. We define the sequence r_t to be the running average of the variational norm of the (functional) error term:

$$r_{t+1} \triangleq (1-\tilde{\eta}_t) r_t + \frac{\tilde{\eta}_t}{1-\kappa}(\|\varepsilon_{\hat{\beta}_t}\|_{\text{var}} + \|\iota_{\hat{\alpha}_t}\|_{\text{var}}).$$

We thus have, for all $t > 0$, $e_t \leqslant r_t$. Using Lemma 3, the sequence $(r_t)_t$ converges towards the scalar expected value

$$r_\infty \triangleq \frac{1}{1-\kappa}\mathbb{E}_{\hat{\alpha},\hat{\beta}}[\|\varepsilon_{\hat{\beta}}\|_{\text{var}} + \|\iota_{\hat{\alpha}}\|_{\text{var}}] > 0. \tag{9}$$

We now relate r_∞ to the number of samples n using a uniform law of large number result on parametric functions. We write $\hat{\beta} = \hat{\beta}_n$ to make explicit the dependency of the quantities on the batch size n.

Using Lemma 2, we bound the quantity

$$\begin{aligned}
E_n \triangleq \mathbb{E}_{\hat{\beta}_n}\|\varepsilon_{\hat{\beta}_n}\|_{\text{var}} &= \mathbb{E}_{\hat{\beta}_n}\|\exp(-T_\beta(g_0^\star)) - \exp(-T_{\hat{\beta}_n}(g_0^\star))\|_\infty \\
&= \mathbb{E}_{Y_1,\dots Y_n\sim\beta} \sup_{x\in\mathcal{X}} \Big| \frac{1}{n}\sum_{i=1}^n \exp(g^\star(Y_i)) - C(x,Y_i)) \\
&\qquad\qquad - \mathbb{E}_{Y\sim\beta}[\exp(g_0^\star(Y)) - C(x,Y)]\Big| \\
&= \mathbb{E}\sup_{x\in\mathcal{X}} |\frac{1}{n}\sum_{i=1}^n \varphi_i(x) - \varphi(x)|,
\end{aligned}$$

where we have defined $\varphi_i : x \to \exp(g^\star(Y_i) - C(x,Y_i))$ and set φ to be the expected value of each φ_i. The compactness of $\mathcal{X}$ ensures that the functions are square integrable and uniformly bounded. Lemma 2 ensures that there exists $S(g^\star)$ such that

$$E_n \leqslant \frac{S(g^\star)}{\sqrt{n}}.$$

We now bound $\mathbb{E}_{\hat{\beta}_n}||\varepsilon_{\hat{\beta}_n}||_{\text{var}}$ using the quantity E_n. First, we observe that $\|_{\text{var}} = g^\star_{\min} < g^\star < 0$, and there exists $C_{\max} > 0$ such that $0 \leqslant C(x,y) \leqslant C_{\max}$ for all $x, y \in \mathcal{X}$, thanks to the Assumption 1.

$$\begin{aligned}
\delta \triangleq \exp(-\|g^\star\|_{\text{var}} - C_{\max}) \leqslant \exp(-T_\beta(g^\star)) \leqslant 1 \\
\exp(-\|g^\star\|_{\text{var}} - C_{\max}) \leqslant \exp(-T_{\hat{\beta}_n}(g^\star)) \leqslant 1,
\end{aligned}$$

where we have used $g^\star = \|g^\star\|_{\text{var}}$. For all $x \in \mathcal{X}$,

$$|\varepsilon_{\hat{\beta}_n}| = |\log\frac{\exp(-T_{\hat{\beta}_n}(g^\star))}{\exp(-T_\beta(g^\star))}| = \Big|\log\big(1 + \frac{\exp(-T_{\hat{\beta}_n}(g^\star)) - \exp(-T_\beta(g^\star))}{\exp(-T_\beta(g^\star))}\big)\Big|. \tag{10}$$

We first obtain an upper-bound independent of n with the first equality in (10):

$$||\varepsilon_{\hat{\beta}_n}||_{\text{var}} \leqslant ||\varepsilon_{\hat{\beta}_n}||_{\infty} \leqslant \|g^\star\|_{\text{var}} + C_{\max}. \tag{11}$$

We now use the second expression in (10): for n large enough, $E_n < \delta$

$$||\varepsilon_{\hat{\beta}_n}||_{\text{var}} \leqslant \max(\log(1+\frac{E_n}{\delta}), -\log(1-\frac{E_n}{\delta})) = -\log(1-\tilde{E}_n), \tag{12}$$

where we have set $\tilde{E}_n \triangleq \frac{E_n}{\delta}$. On the event $\Omega_n = \{\tilde{E}_n \leqslant \frac{1}{2}\}$, a simple calculation gives $-\log(1-\tilde{E}_n) \leqslant (2\log 2)\tilde{E}_n \leqslant 2\tilde{E}_n$. Thanks to Markov inequality, $\mathbb{P}[\tilde{E}_n > \frac{1}{2}] \leqslant 2\mathbb{E}[\tilde{E}_n]$. We then split the expectation over the event Ω_n, and use inequalities (12) and (11) on each conditional expectation:

$$\begin{aligned}
\mathbb{E}||\varepsilon_{\hat{\beta}_n}||_{\text{var}} &= \mathbb{P}\left[\tilde{E}_n \leqslant \frac{1}{2}\right] \mathbb{E}\left[||\varepsilon_{\hat{\beta}_n}||_{\text{var}} \middle| \tilde{E}_n \leqslant \frac{1}{2}\right] \\
&\quad + \mathbb{P}\left[\tilde{E}_n > \frac{1}{2}\right] \mathbb{E}\left[||\varepsilon_{\hat{\beta}_n}||_{\text{var}} \middle| \tilde{E}_n > \frac{1}{2}\right] \\
&\leqslant \frac{2\varphi(\|g^\star\|_{\text{var}} + C_{\max})S(g^\star)}{\sqrt{n}} \\
&\leqslant \frac{4\exp(\|g^\star\|_{\text{var}} + C_{\max})S(g^\star)}{\sqrt{n}} \triangleq \frac{A(g^\star)}{\sqrt{n}}
\end{aligned} \tag{13}$$

The constants S depends on the complexity of estimating the functional $x \to \int_y \exp(g^\star(y) - C(x,y))\mathrm{d}\beta(y)$ with samples from β. A parallel result holds for $\mathbb{E}_{\hat{\alpha}_n}\|\iota_{\hat{\alpha}_n}\|_{\text{var}}$. Therefore, there exists $A(f^\star), A(g^\star) > 0$ such that $r_\infty \leqslant \frac{A(f^\star)+A(g^\star)}{\sqrt{n}}$. As for all $t > 0$, $e_t \leqslant r_t \to_{t\to\infty} r_\infty$, the proposition follows, writing $A = A(f^\star) + A(g^\star)$.

The constant A is larger than $\exp(C_{\max})$ when $C_{\max} \to \infty$; Hence it behaves at least like $\exp(\frac{1}{\varepsilon})$ when $\varepsilon \to 0$.

Note that we have used twice a corollary of the law of large numbers: once when averaging over t with $t \to \infty$ (Eq. (9)), and once when averaging over n with n finite (Eq. (13)). □

A.4 Proof of Proposition 3

In the proof of Proposition 2 and in particular Eq. (8), the term that prevents the convergence of e_t is

$$\eta_t(\|\varepsilon_{\hat{\beta}_t}\|_{\text{var}} + \|\iota_{\hat{\alpha}_t}\|_{\text{var}}),$$

which is not summable in general. We can control this term by increasing the size of $\hat{\alpha}_t$ and $\hat{\beta}_t$ with time, at a sufficient rate: this is what Assumption 3 ensures.

Proof. From Eq. (8), for all $t > 0$, we have

$$0 \leqslant e_{t+1} \leqslant (1-\tilde{\eta}_t)e_t + \eta_t(\|\varepsilon_{\hat{\beta}_t}\|_{\text{var}} + \|\iota_{\hat{\alpha}_t}\|_{\text{var}}). \tag{14}$$

Taking the expectation and using the uniform law of large number (13),

$$\begin{aligned}\mathbb{E}e_{t+1} &\leqslant (1-(1-\kappa)\eta_t)\mathbb{E}e_t + \eta_t \frac{A}{\sqrt{n(t)}} \\ &= (1-(1-\kappa)\eta_t)\mathbb{E}e_t + A\eta_t w_t,\end{aligned} \tag{15}$$

where we have used the definition of $n(t)$ from Assumption 3 in the last line.

The proof follows from a simple asymptotic analysis of the sequence $(\mathbb{E}e_t)_t$, following recursion (15). For all $t > 0$,

$$\mathbb{E}e_{t+1} - \mathbb{E}e_t = -(1-\kappa)\eta_t \mathbb{E}e_t + A\eta_t w_t \leqslant A\eta_t w_t \tag{16}$$

Therefore, from Assumption 3, $(\mathbb{E}e_{t+1} - \mathbb{E}e_t)_t$ is summable and $\mathbb{E}e_t \to_{t\to\infty} \ell \geqslant 0$. Let's assume $\ell > 0$. Summing (16) over t, we obtain

$$\mathbb{E}e_t \leqslant \mathbb{E}e_1 - (1-\kappa)\sum_{s=1}^{t-1} \eta_s \mathbb{E}_s + A\sum_{s=1}^{t-1} \eta_s w_s \to_{t\to\infty} -\infty,$$

which leads to a contradiction. Therefore $\mathbb{E}e_t \to_{t\to\infty} 0$. As $e_t \geqslant 0$ for all $t > 0$, this implies that $e_t \to_{t\to\infty} 0$ almost surely. □

A.5 Proof of Proposition 4

Proof. The proof of Proposition 3 allows us to derive non-asymptotic rates for potential estimations using the online Sinkhorn algorithm. Let us set $\eta_t = \frac{\lambda}{t^a}$, $n(t) = \lceil Bt^{2b} \rceil$ in (14), so that Assumption 3 is met. $\lceil \cdot \rceil$ denotes the ceiling function. We are left to study the recursion (15):

$$\delta_{t+1} \triangleq \mathbb{E}e_{t+1} \leqslant (1 - \frac{\lambda(1-\kappa)}{t^a})\delta_t + \frac{A\lambda}{\sqrt{B}t^{a+b}}$$

Following the derivations of Moulines and Bach (2011, Theorem 2), we have the following bias-variance decomposed upper-bound, provided that $0 \leqslant a < 1$ and $a + b > 1$. For all $t > 0$,

$$\delta_t \leqslant (\delta_0 + \frac{AS}{(a+b-1)\sqrt{B}}) \exp(-\frac{S(1-\kappa)}{2} t^{1-a}) + \frac{2AS}{\sqrt{B}(1-\kappa)t^a}. \tag{17}$$

Let us now relate the iteration number t to the number of seen sample N. By definition

$$n_t = \sum_{s=1}^{t} n(s) \leqslant B\sum_{s=1}^{t} s^{2b} + t \leqslant t + \frac{(t+1)^{2b+1} - 1}{2b+1} \leqslant (2t)^{2b+1}.$$

Therefore, when we have seen N samples, the iteration number is superior to $t(N)$, and the expected error δ_N is of the order of $\delta_{t(N)}$, with

$$t(N) = (N/2)^{\frac{1}{2b+1}}. \tag{18}$$

We write $\delta_N = \delta_{t(N)}$. Replacing (18) in (17) yields

$$\delta_n \leqslant (\delta_0 + \frac{A\lambda}{(a+b-1)\sqrt{B}}) \exp\left(-\frac{\lambda(1-\kappa)}{2}(n/2)^{\frac{1-a}{2b+1}}\right) + \frac{2A\lambda}{\sqrt{B}(1-\kappa)(n/2)^{\frac{a}{2b+1}}}. \tag{19}$$

We note that b and a should be as close to 0 as possible to reduce the bias term, while a should be as close to 1 and b as close to 0 as possible to reduce the variance term. Of course, b should remain larger than $1-a$ to ensure convergence.

To obtain the best asymptotical rates (the error is always dominated by the variance term), we set $a = 1-\iota$, $b = 2\iota$, with $\iota \gtrsim 0$. This yields

$$\begin{aligned}\delta_n &\leqslant (\delta_0 + \frac{A\lambda}{\iota\sqrt{B}}) \exp\left(-\frac{\lambda(1-\kappa)}{2}(n/2)^{\frac{\iota}{1+4\iota}}\right) + \frac{2A\lambda}{\sqrt{B}(1-\kappa)(n/2)^{\frac{1-\iota}{1+4\iota}}} \\ &= \mathcal{O}(n^{-\frac{1-\iota}{1+4\iota}}).\end{aligned}$$

This rate is as close to the rate $\mathcal{O}(\frac{1}{n})$ as desired. We may then perform a last soft C-transform (using the n_t seen samples) over the estimated $f_{t(n)}, g_{t(n)}$ to obtain a estimated solution of the dual optimisation problem (2). The Sinkhorn potentials can therefore be estimated with *fast rates*. Note that the upper bound explodes when $\varepsilon \to 0$, as $C_{\max} \to \infty$, hence $A \to \infty$, and $(1-\kappa) \to 0$. □

Estimating the Sinkhorn distance. The Sinkhorn distance requires to estimate the integral

$$\mathcal{W}(\alpha, \beta) = \int_x f^\star(x) \mathrm{d}\alpha(x) + \int_y g^\star(y) \mathrm{d}\beta(y).$$

At iteration $t(n)$, with empirical realization $\bar{\alpha}_t$ and $\bar{\beta}_t$, containing n samples, we use the estimator

$$\hat{\mathcal{W}}(\alpha, \beta) = \frac{1}{n}\sum_{i=1}^n f_{t(n)}(x_i) + \frac{1}{n}\sum_{i=1}^n g_{t(n)}(y_i),$$

We can bound the estimation error $|\hat{\mathcal{W}}(\alpha,\beta) - \mathcal{W}(\alpha,\beta)| = \mathcal{O}(\frac{1}{\sqrt{n}})$, dominated by the integral evaluation noise. We thus recover a new estimator of the Sinkhorn distance with the same sample complexity as the batch Sinkhorn estimator (Genevay, Chizat, et al., 2019). Our estimator enjoys an original rate for estimating the potentials in $\|\cdot\|_{\mathrm{var}}$.

Algorithm 2 Fully-corrective online Sinkhorn

Input: Distribution α and β, learning weights $(\eta_t)_t$ and batch-sizes $(n(t))_t$.
Set $p_{i,1} = q_{i,1} = 0$ for $i \in (0, n_1]$
for $t = 0, \dots, T-1$ **do**
  Sample $(x_i)_{(n_t, n_{t+1}]} \sim \alpha$, $(y_j)_{(n_t, n_{t+1}]} \sim \beta$.
  Evaluate $(\hat{f}_t(x_i))_{i=(0,n_{t+1}]}$, $(\hat{g}_t(y_i))_{i=(0,n_{t+1}]}$ using $(q_{i,t}, p_{i,t}, x_i, y_i)_{i=(0,n_t]}$ in (7).
  $q_{(0,n_{t+1}],t+1} \leftarrow \log \frac{1}{n} + (\hat{g}_t(y_i))_{(0,n_{t+1}]}$, $p_{(n_t,n_{t+1}],t+1} \leftarrow \log \frac{1}{n} + (\hat{f}_t(x_i))_{(n_t,n_{t+1}]}$.
Returns: $\hat{f}_T : (q_{i,T}, y_i)_{(0,n_T]}$ and $\hat{g}_T : (p_{i,T}, x_i)_{(0,n_T]}$

Algorithm 3 Online Sinkhorn potentials in the discrete setting

Input: Distribution $\alpha \in \triangle^N$ and $\beta \in \triangle^N$, $x \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^{n \times d}$, learning weights $(\eta_t)_t$
Set $p = q = -\infty \in \mathbb{R}^n$.
for $t = 1, \dots, T$ **do**
  $q \leftarrow q + \log(1 - \eta_t)$, $p \leftarrow p + \log(1 - \eta_t)$.
  Sample $J_t \subset [1, N]$, $I_t \subset [1, N]$ of size $n(t)$.
  for $i \in J_t$ **do**
    $q_i \leftarrow \log\Big(\exp(q_i) + \exp\big(\log(\eta_t) - \log \frac{1}{N} \sum_{j=1}^N \exp(p_j - C(x_j, y_i))\big)\Big)$.
  for $i \in I_t$ **do**
    $p_i \leftarrow \log\Big(\exp(q_i) + \exp\big(\log(\eta_t) - \log \frac{1}{M} \sum_{j=1}^M \exp(q_j - C(x_i, y_j))\big)\Big)$.
Returns $f_T : (q, y)$ and $g_T : (p, x)$

B Online Sinkhorn variants

B.1 Fully-corrective scheme

We report the fully-corrective online Sinkhorn algorithm in Algorithm 2. This algorithm also enjoys almost sure convergence, provided that the following assumption is met.

Assumption 4. *For all $t > 0$, the total batch-size $n_t = \frac{B}{w_t^2}$ is an integer. The step-size η_t and the batch-size n_t grows so that $\sum w_t \eta_t < \infty$ and $\sum \eta_t = \infty$.*

With full correction, the total number of observed samples n_t needs to grow at the same rate as the single-iteration batch-size $n(t)$ in Assumption 3. For $\eta_t = \frac{1}{t^a}$, $a \in (1/2, 1]$, it is sufficient to use a constant batch-size $n(t) = B$ to meet Assumption 4. We then have the following property

Proposition 6. *Under Assumption 1 and 4, the fully-corrective online Sinkhorn algorithm converges almost surely:*

$$\|\hat{f}_t - f^\star\|_{\text{var}} + \|\hat{g}_t - g^\star\|_{\text{var}} \to 0.$$

Proof. Using the fully-corrective scheme allows to replace $n(t)$ by $n_t = \sum_{s=0}^{t} n(s)$ in (15). The proposition is then obtained in the same way as Proposition 4. □

B.2 Online Sinkhorn for discrete distributions

The online Sinkhorn algorithm takes a simpler form with discrete distributions. We derive it in Algorithm 3. We set α and β to have size N and M, respectively. We evaluate the potentials as

$$g_t(y) = -\log \sum_{j=1}^{N} \exp(p_j - C(x_j, y))$$
$$f_t(x) = -\log \sum_{j=1}^{M} \exp(q_j - C(x, y_j)),$$

where $(p_j)_{J\in[1,N]}$ and $(q_j)_{J\in[1,M]}$ are fixed-size vectors. Note that the computations written in Algorithm 3 are in log-space,as they should be implemented to prevent numerical overflows. The sets $|I|$ and $|J|$ can have varying sizes along the algorithm, which allows for example to speed-up the initial Sinkhorn iteration (§5.2). In this case, the cost matrix $\hat{C} = C(x_i, y_j))_{i,j}$ should be progressively recorded along the algorithm iterations.

C Extra numerical experiments

We display and describe the supplementary figures mentionned in the main text, as well as experimental details useful for reproduction.

C.1 Online Sinkhorn and variants

Grids and details for §5.1. We set $(\eta_t, n(t)) = \left(\frac{1}{(1+0.1t)^a}, 100(1+0.1t)^b\right)$, with $(a,b) = (0,2)$, $(a,b) = (\frac{1}{2}, 1)$ and $(a, b = 1, 0)$ (constant batch-sizes). Batch Sinkhorn algorithms uses $N = 100, 1000, 10000$. We train Sinkhorn on $t = 5000$ iterations, and train online Sinkhorn long enough to match the number of computations of the large Sinkhorn reference.

All OS convergence curves. To complete Fig. 1, Fig. 4 report the performance of online Sinkhorn for $\varepsilon \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}]\}$. The comparison of performance remains similar to the one produced in the main text.

Fully-corrective online Sinkhorn. Fig. 5 reports the performance of fully-corrected online Sinkhorn (FCOS). We observe that the fully-corrective scheme is less noisy than the non-corrected one. It is less efficient than OS on low-dimensional problems, but faster on the 10 dimensional problem. For GMM-10D, it outperforms the batch Sinkhorn algorithm with $N = 100, 1000$. Note that we interrupt FCOS for $n_t > 20,000$, as our implementation of the C-transform has a quadratic memory cost in n_t—this cost can be reduced to a linear cost with more careful implementation [1].

Randomized Sinkhorn. Fig. 6 reports the performance of randomized Sinkhorn. In low dimension, randomized Sinkhorn is a reasonable alternative to batch Sinkhorn, as it often outperforms it on average, for the same memory complexity (compare purple to orange curve for instance). In high dimension, batch Sinkhorn tend to perform slightly better.

C.2 OT between Gaussians

We measure the performance of online Sinkhorn to transport one Gaussian distribution α to another β. The potentials $f^\star, g^\star$ are known exactly for this problem, which allows to have a strong golden standard. More precisely, adapting the formulae from Janati et al., 2020, assuming $\alpha \sim \mathcal{N}(\mu, A)$ and $\beta \sim \mathcal{N}(\nu, \beta)$

[1] Using e.g. https://www.kernel-operations.io/keops/index.html

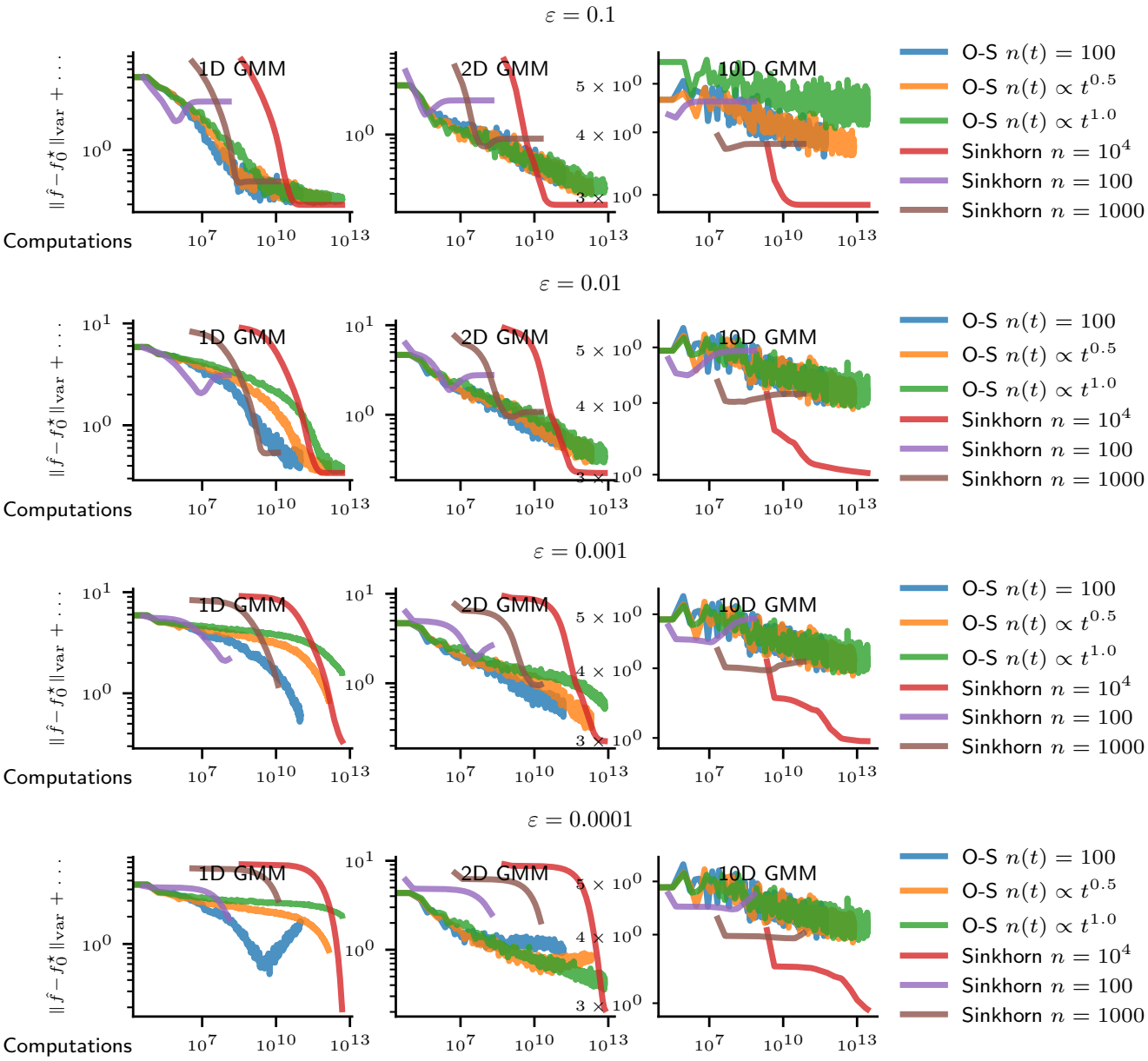


Figure 4: Performance of online Sinkhorn for various ε.

and writing I the identity matrix in $\mathbb{R}^d$, we have

$$C \triangleq (AB + \frac{\varepsilon^2}{4} I)^{1/2}, \quad U \triangleq B(C + \frac{\varepsilon}{2} I)^{-1} - I, \quad V \triangleq A(C + \frac{\varepsilon}{2} I)^{-1} - I$$

$$f^\star : x \to -\frac{1}{2}(x - \mu)^\top U (x - \mu) + x^\top (\mu - \nu)$$

$$g^\star : y \to -\frac{1}{2}(y - \nu)^\top V (y - \nu) + y^\top (\nu - \mu)$$

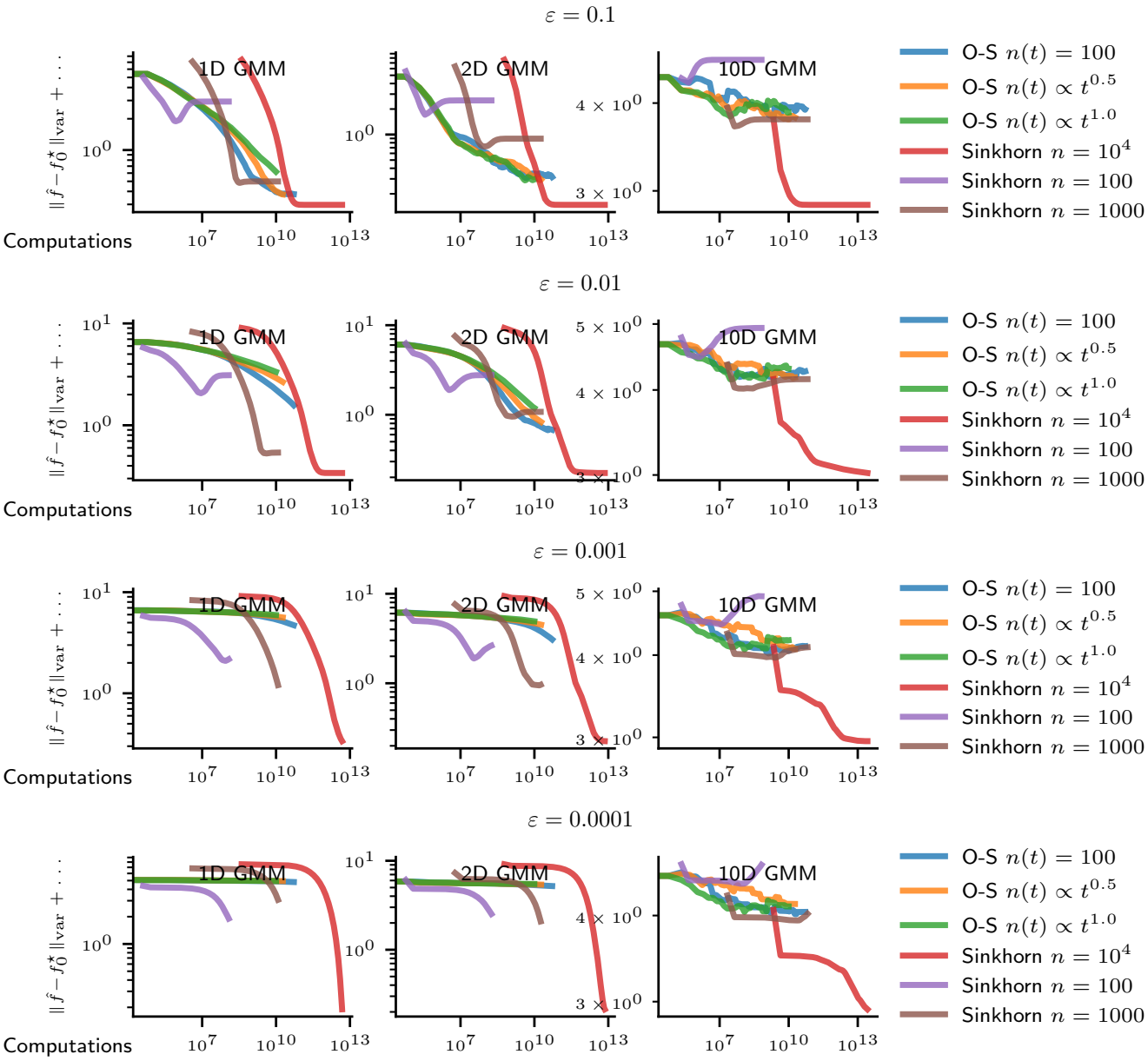


Figure 5: Performance of fully-corrective online Sinkhorn (O-S) for various ε.

We compare batch Sinkhorn ($N = 100, 1000, 10000$) to (non fully-corrected) online Sinkhorn, with $n(t) = B$, and $n(t) = B(1 + 0.1t)^{1/2}$, $B = 100$, and $\varepsilon \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$.

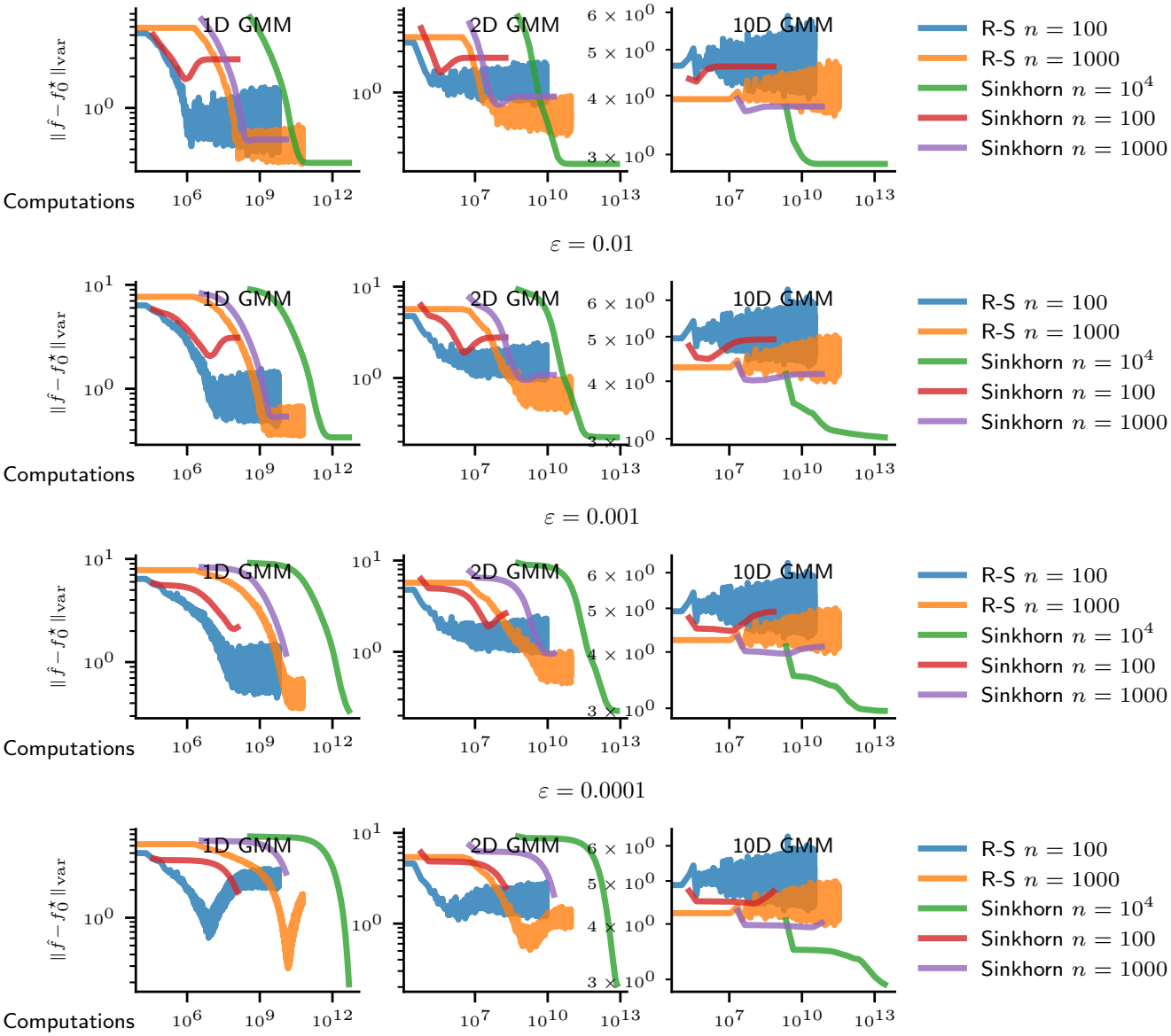


Figure 6: Performance of randomized Sinkhorn (R-S) for various ε.

Results. As displayed in Fig. 7, online Sinkhorn outperforms batch Sinkhorn for all tested batch sizes and all ε. It is faster and does not converge towards biased potentials. This suggests that the performance of online Sinkhorn may be underestimated in the previous analyses due to poor potential reference.

C.3 Illustration of online Sinkhorn potentials on a 2D GMM

The estimate $\hat{f}_t$ is useful to compute the gradient of the Sinkhorn distance $\mathcal{W}(\alpha, \beta)$ with respect to the distribution α. This is useful when α is a parametric distribution α_θ, as it allows to compute the gradient of the Sinkhorn distance with respect to θ using backpropagation. For simplicity, let us assume that $\alpha = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. Then, for all $i \in [1, n]$,

$$\frac{\partial \mathcal{W}(\alpha, \beta)}{\partial x_i} = \nabla_x\big(x \to f^\star(\alpha, \beta)\big)(x_i),$$

so that $\nabla_x f^\star(\alpha, \beta)$ provides a *displacement field* that can be descended to minimize $\alpha \to \mathcal{W}(\alpha, \beta)$. Such point of view can be extended to general distributions using the mean-field point of view, see e.g. Chizat, 2019; Santambrogio, 2015. Estimating $\nabla_x f^\star(\alpha, \beta)$ is therefore crucial to train e.g. generator networks. Both the online Sinkhorn and the batch Sinkhorn algorithm allow to estimate this vector field, through the plug-in estimator $x \to \nabla_x \hat{f}_t$, easily computed using the form (7) of $\hat{f}_t$.

Experiment. With 2D GMMs, we estimate a reference vector field $\nabla f_0^\star$ using Sinkhorn on $N = 10,000$ samples and qualitatively compare the estimations provided by online Sinkhorn and batch Sinkhorn ($N = 1,000$), for the same number of computations.

Results. We represent the estimations $\nabla_x \hat{f}_t$ in Fig. 8, for 10^8 computations. We compare them to a reference displacement field, estimated wityh 10^10 computations. We observe that online Sinkhorn estimates a smoother displacement field than batch Sinkhorn for the same computational budget, that is closer to the reference displacement field. In particular, it is less noisy in low-mass areas. This suggest that online Sinkhorn would be a interesting replacement for batch Sinkhorn in training generative architectures (used by e.g. Genevay, Peyré, et al. (2018)). α_θ is then defined as the push-forward of some simple measure with a neural network g_θ. We leave this direction for future work.

C.4 Online Sinkhorn as a warmup process

Grids and details for §5.2. We set $(\eta_t, n(t)) = \big(\frac{1}{(1+0.1t)^a}, 100(1+0.1t)^b\big)$, with $(a, b) = (0, 2)$, $(a, b) = (\frac{1}{2}, 1)$ and $(a, b = 1, 0)$ (constant batch-sizes). The batch Sinkhorn algorithm that is used for reference and after warmup uses $N = 10000$. In the reference algorithm, we precompute the distance matrix to save computation. In the warmup algorithm, this distance matrix is filled progressively and then kept in memory to perform C-transforms.

We evaluated OS and fully-corrective OS, and found that fully-corrective was less efficient (due to its higher cost in the early iterations). We evaluated sampling with and without replacement in the warmup phase, and found sampling without replacement to be more efficient.

All warmup convergence curves. To complete Fig. 3, we report convergence curves for different ε in Fig. 9. We find that speed-up increased with ε and both the 2D and 3D problems, but remains limited for the 10D problem.

D Stochastic mirror descent interpretation

The online Sinkhorn can be understood as a stochastic mirror descent algorithm for a non-convex problem. This equivalence is obtained by applying a change of variable in (1), defining

$$\mu \triangleq \alpha \exp(f) \quad \text{and} \quad \nu \triangleq \beta \exp(g). \tag{20}$$

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

$$- \min_{(\mu,\nu)\in\mathcal{M}^+(\mathcal{X})^2} \mathrm{KL}(\alpha|\mu) + \mathrm{KL}(\beta|\nu) + \langle \mu \otimes \nu,\, e^{-C} \rangle - 1, \tag{21}$$

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 and 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 (21) 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 7. *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* (20).

Interpretation. It is important to realize that μ_t and ν_t do not need to be stored in memory. Instead, their associated potentials f_t and g_t are parametrized as (7). In particular, μ_t and ν_t remain 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, as in Hsieh et al. (2018).

Finally, we mention that when computing exact gradients (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 variables. 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) and on a semi-dual formulation (Léger, 2019).

Note that one can not directly apply the proofs of convergence of mirror descent to our problem, as the lack of convexity of problem (21) prevents their use.

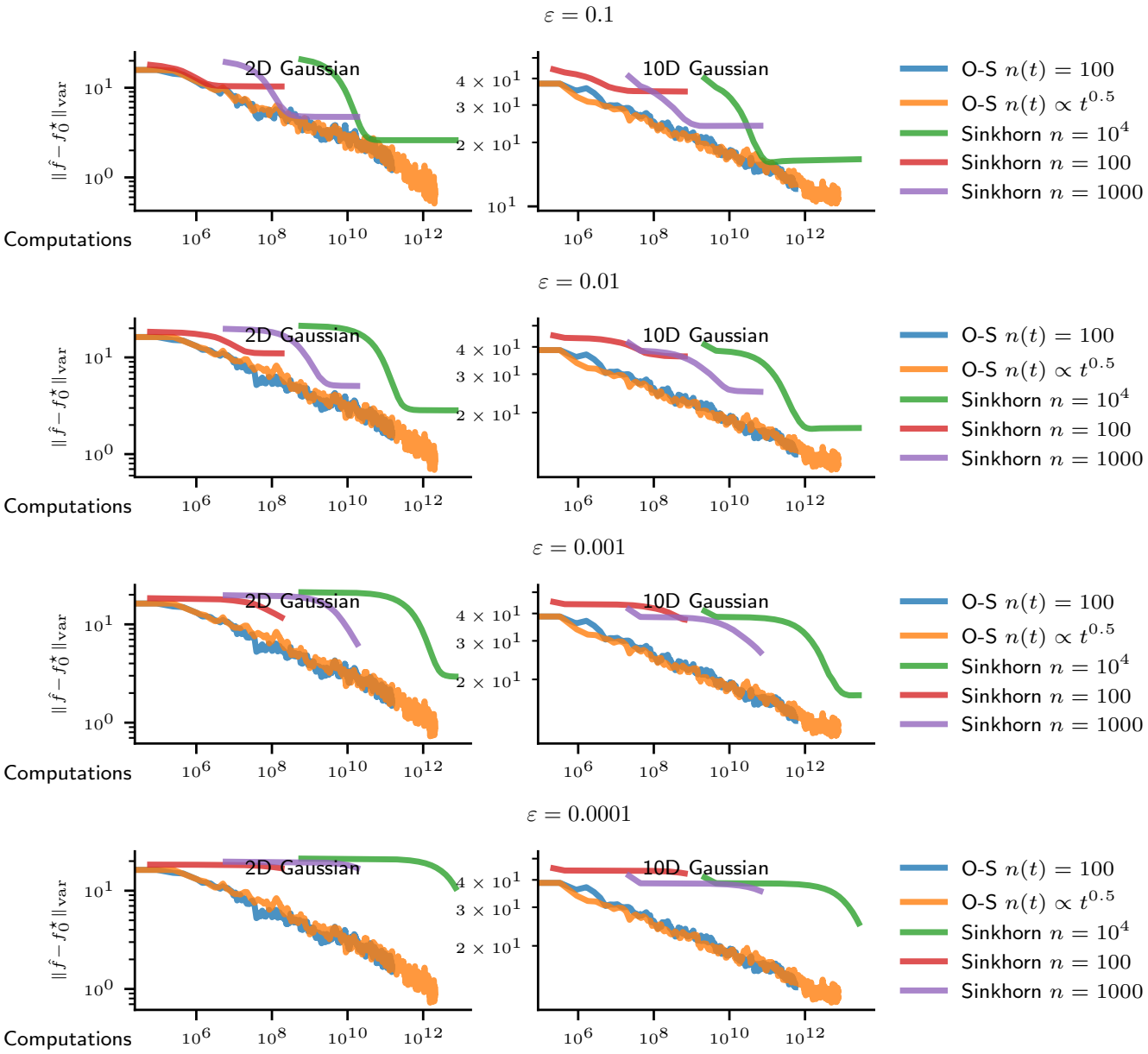


Figure 7: Performance of online-Sinkhorn to estimate OT between two Gaussians. Online Sinkhorn systematically outperforms batch Sinkhorn, but in term of speed and correction.

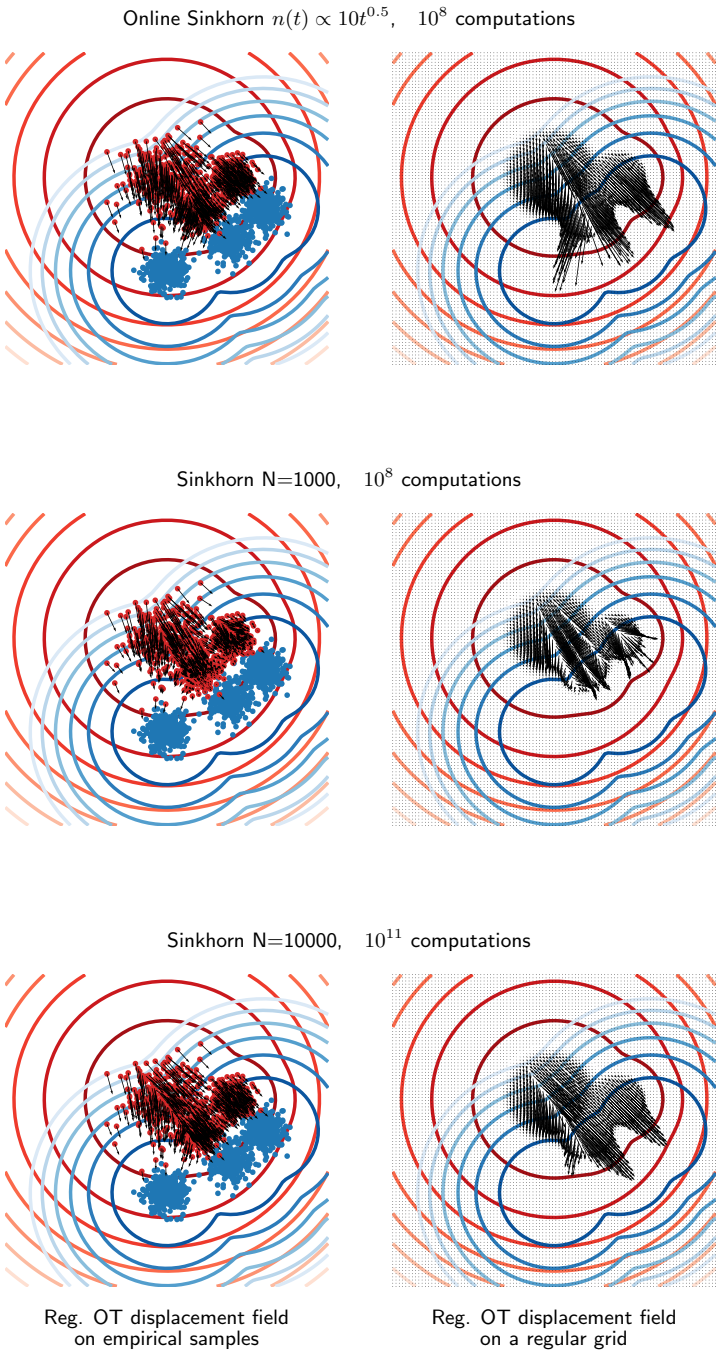


Figure 8: Displacement field as defined by the potentials estimated by online-Sinkhorn and Sinkhorn on a 2D GMM. With the same computational budget, online Sinkhorn finds smoother displacement fields than Sinkhorn. Those are closer to the true reference displacement field (we use Sinkhorn on $N = 10000$ to estimate this reference). α and β log-likelihood level-lines are displayed in red and blue, while the arrows are proportional to $\nabla_x \hat{f}_t(x) \mathrm{d}\alpha(x)$.

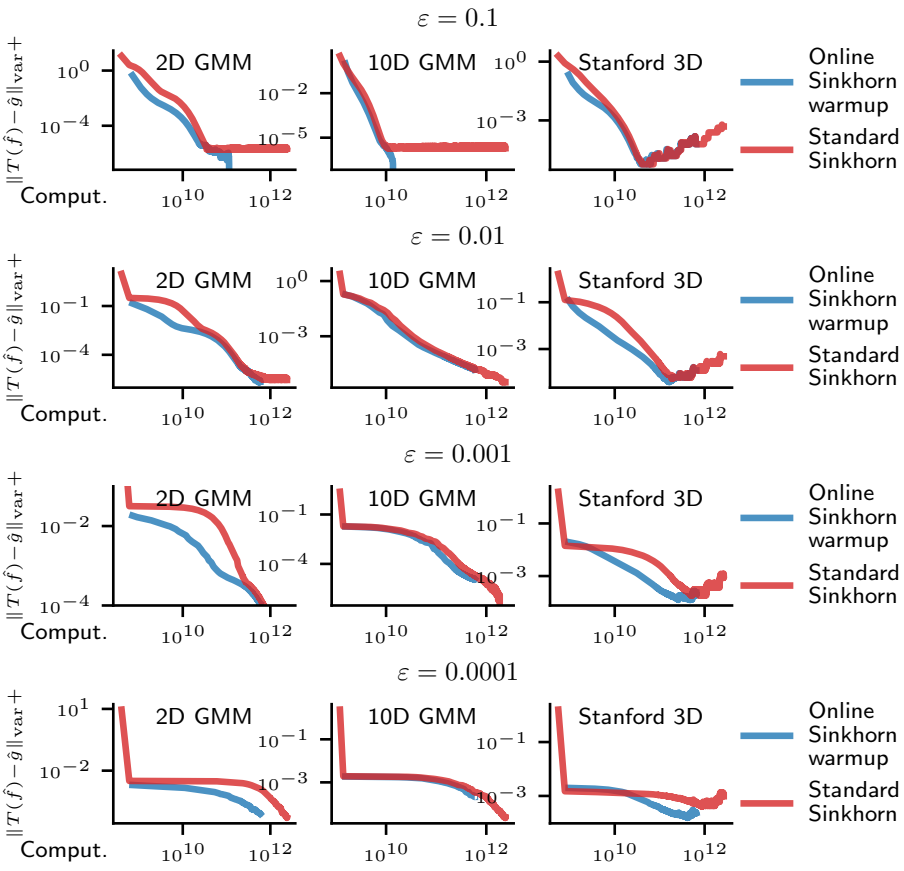


Figure 9: Performance of online-Sinkhorn as warmup for various ε.