Improving Approximate Optimal Transport Distances using Quantization

Gaspard Beugnot[1, 2] **Aude Genevay**[1] **Kristjan Greenewald**[3] **Justin Solomon**[1]

[1]MIT CSAIL
[2]INRIA
[3]IBM Watson AI Lab

Abstract

Optimal transport (OT) is a popular tool in machine learning to compare probability measures geometrically, but it comes with substantial computational burden. Linear programming algorithms for computing OT distances scale cubically in the size of the input, making OT impractical in the large-sample regime. We introduce a practical algorithm, which relies on a quantization step, to estimate OT distances between measures given cheap sample access. We also provide a variant of our algorithm to improve the performance of approximate solvers, focusing on those for entropy-regularized transport. We give theoretical guarantees on the benefits of this quantization step and display experiments showing that it behaves well in practice, providing a practical approximation algorithm that can be used as a drop-in replacement for existing OT estimators.

Optimal transport (OT) is a versatile component of the probabilistic toolbox for machine learning. As an alternative to conventional divergences between probability measures, OT provides a means of measuring how distributions align geometrically. OT has found application in parameter estimation [Bernton et al., 2019], robust learning [Esfahani and Kuhn, 2018], and generative modeling [Salimans et al., 2018, Genevay et al., 2018]—among other learning tasks.

When distributions are absolutely continuous or composed of huge numbers of points, it becomes infeasible to compute OT distances exactly. In this setting, a common approximation follows two steps: First, we draw k samples from both distributions, and then we use linear programming to extract the distance between empirical distributions. This plug-in procedure produces a convergent approximation as $k \to \infty$ (by the Glivenko–Cantelli theorem, since the Wasserstein distance metrizes weak convergence [Villani, 2003]), but two challenges conspire to limit its scalability:

- *Sample complexity* bounds and related results show that this approximation converges with rate $k^{-1/d}$, where d is the ambient dimension [Dudley, 1969, Weed and Bach, 2019]. These sharp *asymptotic* rates exhibit a curse of dimensionality: we need a large number k of samples (growing exponentially with d) before the approximation is useful.
- The *computational complexity* of solving the linear program is roughly cubic in k [Burkard et al., 2012], limiting the maximum k we can take before this method becomes unreasonably slow.

Together, these facts imply that the largest k for which solving the linear program is feasible may not be sufficient for extracting a usable distance estimate, i.e., the bottleneck is not availability of samples/data (the classic statistical setting), but *computation* budget.

Our work is motivated by a simple observation about the methodology above. In machine learning, it is often straightforward to sample from the input measures for OT, e.g. when they come from large datasets, generative models, or easily-sampled smooth distributions. In this case, limited approximation quality is a byproduct of the cubic computational expense rather than a paucity of samples. The algorithm above only draws $O(k)$ samples—but it could draw more without affecting the asymptotic runtime. That is, we can improve approximation quality with little added computational expense by drawing more than k samples, cutting down to k *representative* (weighted) samples, and then solving a smaller discrete problem.

We introduce a practical, easily-implemented improvement to empirical OT. In our algorithm, the OT solver remains either the linear program solver or the recently-popular regularized Sinkhorn algorithm [Cuturi, 2013]. As input to this step, however, we "summarize" a superlinear number of samples with k weighted samples through quantization. Our technique is seamless to implement given an implementation of empirical OT and substantially improves approximation quality given fixed computational cost. It can be used as a

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:290–300.

drop-in replacement for existing estimators. Beyond verifying performance empirically, we provide theory predicting the behavior we observe, in the low quantization error setting. While it is impossible to overcome the asymptotic curse of dimensionality associated to all finitely-supported measures [Kloeckner, 2012], our method leverages better convergence rates in the finite sample regime for "clusterable" distributions [Weed and Bach, 2019]. This leads to substantial practical benefit, with an improvement of the exponent of the convergence rate by a factor 2 in the best case (fast decaying tails) or at worst on par with the plug-in estimator (close to uniform).

Related work. OT suffers from a severe curse of dimensionality. Effective approximation requires an exponential number of samples n in the ambient dimension. For an absolutely continuous measure μ (w.r.t. Lebesgue), its Wasserstein distance to any measure supported on n points is asymptotically lower-bounded by $O(n^{-1/d})$ [Dudley, 1969]. This bound can sometimes be circumvented, e.g., when the measures have lower intrinsic dimension [Weed and Bach, 2019] or when the support is discrete (convergence rate $O(\sqrt{1/n})$, with constant depending on dimension) [Sommerfeld et al., 2018]. To counter this curse of dimensionality, the best-known workaround relies on entropic regularization, with $O(\sqrt{1/n})$ convergence [Genevay et al., 2019]. Another estimator penalizes the rank of the transport plan [Forrow et al., 2019], while [Goldfeld and Greenewald, 2020] proposes a smoothed distance by convolving measures with Gaussians. While these exhibit better convergence rates, they only approximate the Wasserstein distance and do not converge to its true value. The curse of dimensionality can sometimes be mitigated for standard OT—[Weed and Bach, 2019] proves that for mixtures of Gaussians and clusterable distributions, the p-th power of the p-Wasserstein distance enjoys a $O(\sqrt{1/n})$ rate for small n—implying a $O(n^{-1/4})$ rate for W_2.

While the curse of dimensionality requires many samples to approximate transport reliably, in practice computational complexity prevents us from doing so. OT between discrete measures yields a large-scale linear program solvable using network flow solvers or the Hungarian algorithm, when both measures have the same size and uniform weights [Burkard et al., 2012]. These take $O(n^3 \log n)$ time, where n is the support size. As a faster alternative, entropy-regularized OT can be solved with quadratic complexity using Sinkhorn's algorithm [Sinkhorn, 1967], but its convergence rate decays when regularization goes to zero [Franklin and Lorenz, 1989].

For efficient OT approximation, we oversample the input measures and compute a summary via a quantization algorithm like k-means; note quantization is equivalent to finding the closest measure supported on k points in 2-Wasserstein distance [Pollard, 1982, Canas and Rosasco, 2012]. The original k-means algorithm [Lloyd, 1982] is prohibitive for large sample sizes and often reaches local minima. With a careful initialization, however, [Arthur and Vassilvitskii, 2006] proved that k-means likely converges to near its global optimum. This initialization, called k-means++, is obtained via D^2 sampling and is $O(\log k)$-close to optimal in expectation. This yields a cheap approximation in $O(nk)$ time, since the algorithm requires k passes through the data. Later variants have lower computational complexity, among which [Bahmani et al., 2012] performs only a fixed number of passes on the data and [Bachem et al., 2016] uses an MCMC D^2 sampler. These benefit from bounds similar to k-means++ but have $O(n)$ computational complexity.

Our approach has similarities with a line of work that uses a multi-scale scheme to compute optimal transport efficiently [Schmitzer and Schnörr, 2013, Gerber and Maggioni, 2017]. However, they focus on accelerating the exact computation of optimal transport, while we target a fast approximation. These multi-scale approaches also do not leverage a connection between k-means and optimal transport to yield quantitative analysis, and they are not applicable to entropy-regularized transport.

Contributions. We propose efficient OT estimators using quantization, with theoretical analysis for two classes of OT problems:

- **(Unregularized) OT:** We leverage the link between OT and k-means [Pollard, 1982, Canas and Rosasco, 2012] to quantify the bias and give precise bounds for Gaussian mixtures and clusterable distributions in the non-asymptotic regime.
- **Entropy-regularized OT**: Building on complexity results for Sinkhorn [Altschuler et al., 2017], we prove that our pre-processing can yield ε-approximate OT with better time/space complexity.

We compare our estimators to the plug-in estimator on toy and real-world datasets.

Notation. Let μ and ν be probability measures on a compact set $\mathscr{X} \subseteq \mathbb{R}^d$. The 2-Wasserstein distance between μ and ν is

$$W_2(\mu,\nu) \stackrel{\text{def.}}{=} \left(\min_{\pi \in \Pi(\mu,\nu)} \int_{\mathscr{X}\times\mathscr{X}} \|x-y\|_2^2 \, d\pi(x,y) \right)^{1/2}, \quad (1)$$

where $\Pi(\mu,\nu)$ is the set of couplings on $\mathscr{X} \times \mathscr{X}$ with marginals μ, ν. Given n samples from each measure, $X_n \stackrel{\text{def.}}{=} (x_1,\dots,x_n) \sim \mu^{\otimes n}$ and $Y_n \stackrel{\text{def.}}{=} (y_1,\dots,y_n) \sim \nu^{\otimes n}$, the *empirical plug-in estimator* for W_2 is

$$W_2(\hat{\alpha}_n, \hat{\beta}_n) = \left(\min_{\substack{\pi \mathbb{1} = \mathbb{1}/n \\ \pi^T \mathbb{1} = \mathbb{1}/n}} \frac{1}{n^2} \sum_{i,j=1}^{n} \|x_i - y_j\|_2^2 \pi_{ij} \right)^{1/2}, \quad (2)$$

where $\hat{\alpha}_n \stackrel{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\hat{\beta}_n \stackrel{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ are empirical measures from μ and ν, resp.

1 ALGORITHM OVERVIEW

We aim to improve the plug-in estimator $W_2(\hat{\mu}_k, \hat{\nu}_k)$, which approximates $W_2(\mu,\nu)$ with $O(k^3 \log k)$ computational complexity (that of LP solvers) and $O(k^{-\alpha})$ bias, given k samples from each measure. In the worst case (e.g., uniform distributions), $\alpha = 1/d$, but there exist regimes in which the rate improves (see §2.2). Our idea is to *oversample* the measures, using $n > k$ samples to construct approximations of μ and ν of size k that yield an estimated OT value with better bias while preserving computational complexity. To satisfy these criteria, we need to ensure that pre-processing takes $O(k^3 \log k)$ time.

We denote by $\hat{S}_k(X_n)$ a stochastic map that inputs a sample $X_n = (x_1,\ldots,x_n) \sim \mu^{\otimes n}$ and outputs a k-point quantization. For any finite $S \subseteq \mathscr{X}$, use the function $P_S : \mathscr{X} \to S$ to denote the function that maps any point in $\mathscr{X}$ to its nearest neighbor in S. Denoting by $\hat{\mu}_n$ (resp., $\hat{\nu}_n$) the empirical measure associated to the n-sample X_n (resp., Y_n) and $f_{\#}(\mu)$ the pushforward of μ through f, our estimator is defined as:

$$\mathrm{Est}(k,n) \overset{\text{def.}}{=} W_2(P_{\hat{S}_k(X_n)\#}(\hat{\mu}_n), P_{\hat{S}_k(Y_n)\#}(\hat{\nu}_n)). \tag{3}$$

That is, we replace $\hat{\mu}_k$, $\hat{\nu}_k$ in the plug-in estimator (2) with weighted k-point measures $P_{\hat{S}_k(X_n)\#}(\hat{\mu}_n)$ and $P_{\hat{S}_k(Y_n)\#}(\hat{\nu}_n)$, the centers of approximate k-means on X_n and Y_n, resp. Each center is weighted proportionally to the number of samples in its Voronoi region. The plug-in estimator (2) corresponds to $n = k$.

There are two steps in our pre-processing: *(i)* selecting k points representative of the larger n samples and *(ii)* weighting the resulting k points with the number of samples in their Voronoi regions. For k-means++, *(i)* is $O(nk)$ while for [Bachem et al., 2016, Bahmani et al., 2012] it is $O(k)$. Regardless, the assignment in step *(ii)* requires $O(nk)$ time. To be consistent with the $O(k^3 \log k)$ time complexity of the OT solver, we thus set $n = k^2 \log k$.

Algorithm 1 summarizes our estimator. It takes four steps: (1) sample $k^2 \log k$ points from each measure, (2) run k-means++ initialization, (3) project the $k^2 \log k$ points onto the k cluster centers, and (4) compute OT between these new weighted point clouds. Steps (1) and (3) are seamless to implement, while steps (2) and (4) have readily available implementations in many languages, as they come from well-known algorithms. Thus, the procedure is highly practical, and it can easily be implemented to improve the bias of OT estimation with similar running times.

The performance of our approach is summarized informally in the theorem below; we bound bias in §2.

Theorem 1 (Informal)**.** *Algorithm 1 runs in $O(k^3 \log k)$ time.*[1] *The estimator has $O(k^{-2\alpha})$ bias in the best case and $O(k^{-\alpha})$ at worst, where the latter is the bias of the empirical plug-in estimator.*

Algorithm 1: Approximation of $W_2(\mu,\nu)$

Input : Two samplers μ,ν; number of anchor points k.
Output : Approximation of $W_2(\mu,\nu)$ with complexity $O(k^3 \log k)$

`/* Sample n points */`
Set $n = k^2 \log k$
Sample $X_n = (x_1,\ldots,x_n)$ i.i.d. from μ and $Y_n = (y_1,\ldots,y_n)$ i.i.d. from ν
`/* Subsample k anchor points */`
Compute $\hat{S}_k(X_n) = (c_1,\ldots,c_k)$ with k-means++
Compute $\hat{S}_k(Y_n) = (d_1,\ldots,d_k)$ with k-means++
`/* Compute weights */`
Set $a_i = \sum_{j=1}^n \mathbf{1}_{i=\arg\min_l \|x_j - c_l\|_2^2} \; \forall i \in \{1,\ldots,k\}$
Set $b_i = \sum_{j=1}^n \mathbf{1}_{i=\arg\min_l \|x_j - d_l\|_2^2} \; \forall i \in \{1,\ldots,k\}$
`/* Cost matrix */`
Set $C_{ij} = \|c_i - d_j\|_2^2 \; \forall i,j \in \{1,\ldots,n\}$
`/* Weighted Wasserstein distance */`
return $W_2(P_{\hat{S}_k(X_n)\#}\hat{\mu}_n, P_{\hat{S}_k(Y_n)\#}\hat{\nu}_n) \overset{\text{def.}}{=} L_C(a,b)^{1/2}$

Remark 1. *The "best case" happens in the finite sample regime, when distributions have low quantization error as defined in §2.2. For near-uniform distributions, we get the asymptotic rate right away and cannot hope to improve on the plug-in estimator.*

This theorem predicts the performance observed in §4. In short, with the same computational complexity, we improve the bias by an exponent of 2 compared to the plug-in estimator. Time complexity is a direct addition of pre-processing and LP solver complexities. The bias bounds, on the other hand, require more work and are the object of the next section.

2 THEORETICAL ANALYSIS

2.1 BOUNDING BIAS

The bias of our estimator $\mathrm{Est}(k,n)$ defined in (3) is

$$\mathrm{Bias}(k,n) = \left| W_2(\mu,\nu) - \mathbb{E}\left[\mathrm{Est}(k,n)\right] \right|. \tag{4}$$

By the triangle inequality on $|\cdot|$ and W_2, we have that

$$\begin{aligned}\mathrm{Bias}(k,n) \leq\; & \mathbb{E}_{X_n,\hat{S}_k}\left[W_2(\mu, P_{\hat{S}_k(X_n)\#}(\hat{\mu}_n))\right] \\ & + \mathbb{E}_{Y_n,\hat{S}_k}\left[W_2(\nu, P_{\hat{S}_k(Y_n)\#}(\hat{\nu}_n))\right],\end{aligned}$$

so bounding bias amounts to controlling the two terms above. This requires some definitions:

[1] This complexity assumes sampling is cheap, i.e., $O(1)$. If drawing samples requires complex operations, the number of points we can sample will be below $n = k^2 \log k$; it is straightforward to adapt to this case.

Definition 1 (Quantization error $\phi_S(C)$)**.** *Let $C \subseteq \mathcal{X}$ be a finite set of n elements. For any $S \subseteq \mathcal{X}$, define the* quantization error *of C w.r.t. S as*

$$\phi_S(C) = \sum_{x \in C} d(x,S)^2,$$

where $d(x,S) = \min_{s \in S} d(x,s)$. For $k \leq n$, denote by ϕ_k^{OPT} the optimal quantization error *for a set of k elements, written $\phi_k^{OPT}(C) = \min_{S \subseteq \mathcal{X}, |S|=k} \phi_S(C)$, and S_k its minimizer.*

We can relate the bias of our estimator to sample complexity and quantization error as follows:

Theorem 2 (Bias of the estimator)**.** *Suppose $\mathbb{E}[W_2(\mu, \hat{\mu}_n)] \leq O(n^{-\alpha})$, where α is the sample complexity rate of μ. Then, for a sample $X_n \sim \mu^{\otimes n}$,*

$$\mathbb{E}_{X_n, \hat{S}_k}\left[W_2(\mu, P_{\hat{S}_k(X_n)\#}\hat{\mu}_n)\right] \leq O\left(n^{-\alpha} + \sqrt{(\log k)/n}\,\mathbb{E}_{X_n}\left[\phi_k^{OPT}(X_n)\right]^{1/2}\right).$$

The sample complexity here is not necessarily the asymptotic rate $\alpha = 1/d$. Rather, we will see in §2.2 that our estimator performs well in the finite sample regime for clusterable distributions, with rate $\alpha = 1/4$.

Proof. By the triangle inequality on W_2, we can decompose into two quantities A and B:

$$\mathbb{E}_{X_n, \hat{S}_k}\left[W_2(\mu, P_{\hat{S}_k(X_n)\#}\hat{\mu}_n)\right] \leq \underbrace{\mathbb{E}_{X_n} W_2(\mu, \hat{\mu}_n)}_{A} + \underbrace{\mathbb{E}_{X_n, \hat{S}_k} W_2(\hat{\mu}_n, P_{\hat{S}_k(Xn)\#}\hat{\mu}_n)}_{B}. \quad (5)$$

- A is the sample complexity rate of the empirical distribution, which we assume to be $O(n^{-\alpha})$.
- B is the error made when projecting the n samples onto k weighted points chosen by k-means++. If $n = k$, it vanishes and we recover the sample complexity of the empirical estimator. Controlling B requires relating Wasserstein distance to the optimal quantization [Canas and Rosasco, 2012].

Denoting $X_n = (x_1, \ldots, x_n) \sim \mu^{\otimes n}$, we write:

$$B = \mathbb{E}_{X_n, \hat{S}_k}\left[\left(\frac{1}{n}\sum_{i=1}^{n} d(x_i, \hat{S}_k)^2\right)^{1/2}\right] \leq \mathbb{E}_{X_n}\left[\left(\frac{8(\log k + 2)}{n}\phi_k^{\mathrm{OPT}}(X_n)\right)^{1/2}\right]. \quad (6)$$

The first equality comes from the equivalence between W_2 and the quantization error [Canas and Rosasco, 2012, Lemma 1]. The second is the k-means++ optimality bound of [Arthur and Vassilvitskii, 2006]. Jensen's inequality completes the proof. Note having the optimal set S_k instead of $\hat{S}_k$ would remove the $\log k$ factor in (6). □

In our algorithm, we take $n = k^2 \log k$ to get the following bias for our estimator:

Corollary 1. *In the setting of Theorem 2, with $n = k^2 \log k$, our estimator (Algorithm 1) satisfies*

$$\mathbb{E}_{X_n, \hat{S}_k}\left[W_2(\mu, P_{\hat{S}_k(X_n)\#}(\hat{\mu}_n))\right] \leq O\left((k^2 \log k)^{-\alpha} + 1/k\,\mathbb{E}_{X_n}\left[\phi_k^{OPT}(X_n)\right]^{1/2}\right).$$

Corollary 1 tells us that at best, our estimator improves the exponent in the bias bound by a factor of 2, going from $O(k^{-\alpha})$ to $O(k^{-2\alpha})$ while keeping computational complexity on par with the empirical plug-in estimator. To benefit from this improvement, we need to ensure quantization error—the second term in the bound—is small enough so that the first dominates.

2.2 CONTROLLING THE QUANTIZATION ERROR

To prove our estimator improves bias, we must make an assumption on the behavior of the quantization error when quantizing an n-sample from μ on k points. Intuitively, the quantization error is small when the measure is well-concentrated. In particular, we can upper bound quantization error for Gaussian mixtures and measures supported on finite numbers of balls.

Remark 2. *We derive improved theoretical rates for these two classes of functions, but our algorithm is better than the plug-in estimator for* any *dataset whose quantization error is smaller than the sample complexity. This is verified by several real-world datasets (Fig. 8, supplement), underscoring the practical significance of our proposed algorithm.*

Definition 2 (Clusterable distribution)**.** *A distribution μ is an (m, σ^2)-Gaussian mixture if it is a mixture of m Gaussian distributions in $\mathbb{R}^d$ and the trace of the covariance matrix of each mixture component is upper-bounded by σ^2. A distribution μ is (m, Δ)-clusterable if* $\mathrm{supp}(\mu)$ *lies in the union of m balls of radius at most Δ.*

By writing down the definition of ϕ_k^{OPT}, it is straightforward to prove that for k $\geq m$, $1/n \cdot \mathbb{E}[\phi_k^{OPT}(X_n)] \leq \sigma^2$ if μ is a (m, σ^2)-Gaussian mixture, and $1/n \cdot \mathbb{E}[\phi_k^{OPT}(X_n)] \leq \Delta^2$ if μ is (m, Δ)-clusterable.

Incidentally, for such measures, better sample complexity rates can be derived [Weed and Bach, 2019]:

Proposition 1 ([Weed and Bach, 2019])**.** *If μ is a (m, σ^2)-Gaussian mixture and $\log \frac{1}{\sigma} \geq 25/8$, then for all $n \leq m(32\sigma^2 \log \frac{1}{\sigma})^{-2}$,*

$$\mathbb{E}[W_2^2(\mu, \hat{\mu}_n)] \leq 84\sqrt{m/n}. \quad (7)$$

The same rate holds for (m,Δ)-clusterable distributions, for all $n \leq m(2\Delta)^{-4}$.

This result can be extended to distributions that are mixtures with fast decaying tails. This improved rate holds *in the small-sample regime*, but asymptotically, the $1/d$ rate returns. This rate is for *squared* W_2, so in our analysis using W_2 this only implies $\alpha = 1/4$ via Jensen's inequality. Thus, these improved rates for W_2 are only relevant in dimension higher than 4.

Further assumptions on σ^2 (resp. Δ) improve the convergence rate of the bias from Theorem 2:

Proposition 2. *If μ is an (m,σ^2)-Gaussian mixture (resp. (m,Δ)-clusterable), then for all $k \geq m$ such that $k^2 \log k \leq m(32\sigma^2 \log \frac{1}{\sigma})^{-2}$ (resp. $k^2 \log k \leq m(2\Delta)^{-4}$) our estimator (Algorithm 1) satisfies*

$$\mathbb{E}[W_2(\mu, P_{\hat{S}_k(X_n)\#}\hat{\mu}_n)] \leq \sqrt{84}\left(\frac{m}{k^2 \log k}\right)^{1/4} + C\sigma\sqrt{\log k},$$

(replacing σ by Δ in the above bound for clusterable distributions), where C is independent of k and σ. If $k^2 \log k \leq m(32\sigma^2 \log \frac{1}{\sigma})^{-2}$ (resp. $k^2 \log k \leq m(2\Delta)^{-4}$), then $\sigma \leq O((\log k)^{-1/4} k^{-1/2})$ (resp. Δ), and the rate becomes $O((\log k)^{1/4} k^{-1/2})$.

Hence, we achieve an $O((\log k)^{1/4} k^{-1/2})$ rate in $O(k^3)$ computation time, compared to the $O(k^{-1/4})$ rate of the empirical estimator. For the range of k we consider, we observe in practice that the assumption $1/n \cdot \phi_k^{OPT} \leq 1/k$ often holds, and hence our bound applies. Due to the curse of dimensionality, however, there is no guarantee for this to hold in the asymptotic case.

Intuition on the finite sample regime. The intuition for the bound of Proposition 1 is not simple. We provide an informal explanation. From a high level, in the small sample regime, we are looking at a coarse scale (e.g. from a distance, Gaussians "look like" Diracs) so the bound behaves like discrete optimal transport, which is $n^{-1/2}$. However when the number of samples grows, we are looking at a fine scale; in this regime, we suffer from the curse of dimensionality. A second piece of intuition is simpler: when you have very few samples, every new sample brings a lot of information, but after a while, the information gain of each new sample diminishes.

3 REGULARIZED TRANSPORT

Quantization can also improve approximate OT solvers, as it introduces negligible error while improving the required runtime and memory storage, at least in the discrete case. We focus on entropic regularization, a popular approximation of OT obtainable in quadratic time with Sinkhorn's algorithm [Cuturi, 2013]. More precisely, the computational complexity to obtain an ε-approximation of the unregularized cost for discrete problems is bounded by $O(k^2\varepsilon^{-2})$, an order of magnitude cheaper than the linear program [Lin et al., 2019]. The oversampling strategy used previously for absolutely continuous measures is irrelevant, however: quantizing n points with k centroids takes at least $O(nk)$ time (because of weight assignment), which exceeds $O(k^2)$ for $n > k$.

Instead, we consider the case where we are given two very large discrete measures as input and rely on quantization to design a more efficient approximation procedure. In this setting, the literature focuses on *complexity bounds*: given two discrete distributions over n points and a target precision ε, the aim is to provide an ε-approximation of unregularized transport with bounded complexity [Altschuler et al., 2017, Dvurechensky et al., 2018, Lin et al., 2019]. Building on this problem formulation, we propose a quantization step with target precision ε as a preprocessing step. Afterwards, any approximate transport solver can be used on the resulting quantized distribution. This provides the same theoretical guarantees and bounded computational complexity as above, with potential computation time improvements. Our algorithm is detailed in Algorithm 2.

Algorithm 2: ε-approximation of $W_2(\mu_n, \nu_n)$

Input : Finite distributions μ_n, ν_n; target precision ε
Output : 3ε-approximation of $W_2(\mu_n, \nu_n)$ with complexity $O(k^2\varepsilon^{-2})$

`/* Quantize the point clouds */`
$S_\varepsilon = \text{QUANTIZE}(\mu_n, \varepsilon)$; $|S_\varepsilon| = k_{\varepsilon,\mu_n}$
$T_\varepsilon = \text{QUANTIZE}(\nu_n, \varepsilon)$; $|T_\varepsilon| = k_{\varepsilon,\nu_n}$
`/* Compute weights and cost matrix */`
Set $a_i = \sum_{j=1}^n w_{\mu,j} \mathbf{1}_{i=\arg\min_l \|x_j - c_l\|_2^2} \ \forall i \in \{1,\ldots,k_{\varepsilon,\mu_n}\}$
Set $b_i = \sum_{j=1}^n w_{\nu,j} \mathbf{1}_{i=\arg\min_l \|y_j - d_l\|_2^2} \ \forall i \in \{1,\ldots,k_{\varepsilon,\nu_n}\}$
Set $C_{ij} = \|c_i - d_j\|_2^2 \ \forall c_i, d_j \in S_\varepsilon \times T_\varepsilon$
`/* Regularized transport solver */`
return $\text{APPROXOT}(C, a, b, \varepsilon)$

Algorithm 2 relies on two subroutines: QUANTIZE and APPROXOT. The former inputs a point cloud μ_n and a tolerance ε and outputs a (sub)set S_ε, which is a quantized version of μ_n. k-means++ can be adapted easily to do this. An example is in Algorithm 3. APPROXOT yields an ε approximation of unregularized transport. The most used one is probably the Sinkhorn algorithm, which has a complexity bounded by $O(k^2\varepsilon^{-2})$; see [Altschuler et al., 2017] for details. This is the one we use in our experiments.

Algorithm 3 is directly adapted from the original k-means++ algorithm. It is guaranteed to finish, as $S_\varepsilon = \mu_n$ is a solution for any ε. Denoting $k_\varepsilon = |S_\varepsilon| \leq n$, we have that the complexity of Algorithm 3 is bounded by $O(nk_\varepsilon)$. Thus, Algorithm 2 has a complexity bounded by $O(nk_\varepsilon + k_\varepsilon^2\varepsilon^{-2}) \lesssim O(n^2\varepsilon^{-2})$. The fact that it outputs a 3ε approximation of OT relies on

Algorithm 3: QUANTIZE

Input : A finite distribution μ_n with support and weights $(x_i, w_i)_{1\le i\le n)}$; target precision ε.
Output : Set S_ε with k_ε elements, s.t. $W_2^2(\mu_n, P_{\hat{S}_\varepsilon\#}\mu_n) = \sum_i w_i d(x_i, S_\varepsilon)^2 < \varepsilon^2$.

$S_\varepsilon \leftarrow x_{\text{RAND}(1,n)}$
$D = (w_i d(x_i, S_\varepsilon)^2)_{1\le i\le n}$
while $\sum_i D_i > \varepsilon^2$ **do**
  $S_\varepsilon \leftarrow x_{\arg\max_i D_i}$
  $D = (w_i d(x_i, S_\varepsilon)^2)_{1\le i\le n}$
return S_ε

Lemma 1 of [Canas and Rosasco, 2012]:

$$\begin{aligned} W_2(\mu,\nu) \le & W_2(P_{\hat{S}_\varepsilon\#}\mu, P_{\hat{T}_\varepsilon\#}\nu) \\ & + W_2(\mu, P_{\hat{S}_\varepsilon\#}\mu) + W_2(\nu, P_{\hat{S}_\varepsilon\#}\nu) \end{aligned} \tag{8}$$

The first term is approximated within ε thanks to APPROXOT, the second/third thanks to Algorithm 3.

Overall, we have two options to obtain a 3ε approximation of $W_2(\mu_n, \nu_n)$:

- Run APPROXOT$(C_n, w_\mu, w_\nu, 3\varepsilon)$ where C_n is the $n\times n$ cost matrix between μ_n and ν_n.
- Run Algorithm 2.

Both have a complexity $\le O(n^2\varepsilon^{-2})$ and provide the same theoretical guarantees; but the latter can provide a significant speed up. We compare both approaches in the next section, measuring CPU-time vs. precision.

Space complexity. While Sinkhorn's algorithm has space complexity of $O(n^2)$, we highlight that alg. 2 has space complexity of $O(n + k_\varepsilon^2)$. Indeed, the QUANTIZE algorithm only needs to keep track of the assignment of every point to their nearest centroid: this is a vector of size n. Thus, for huge datasets where storage is critical, quantization is a natural way to downscale the point cloud while keeping track of the precision loss.

Remark 3. *Some remarks about Algorithm 2:*

- *The bound on the complexity of* APPROXOT *usually involves* $\|C\|_\infty$. *It will be smaller for the cost between centroids, providing additional speedup.*
- *This preprocessing step can be used for any p-Wasserstein distance, by changing the exponent in* QUANTIZE *accordingly* ($D = (w_i d(x_i, S_\varepsilon)^p)_{1\le i\le n}$).
- *We provide an algorithm with the same approximation guarantees than the baseline, with lower or equal computational complexity. A sharp bound on the output of algorithm 2 would require studying* $\varepsilon \mapsto k_\varepsilon$.

4 EXPERIMENTS

Datasets. We test on discrete (mainly real-world data) and continuous (synthetic) distributions. The latter tests theoretical bounds, while the former shows efficiency of Algorithm 1 on large point clouds. Fig. 5 (supplement) shows examples. The discrete datasets are: *DOT*, *Adult*, and *Sampled Mixtures*. The 'true' distance is computed on the whole point cloud; some datasets were downsampled to suit ground truth computation on our machine. *DOT* [Schrieber et al., 2017] contains grayscale images (i.e., fixed discrete support in $\mathbb{R}^2$) in various resolutions, a benchmark used e.g. in [Sommerfeld et al., 2018], which uses the plug-in estimator. *Adult* (UCI repository) is a point cloud in $\mathbb{R}^6$ with continuous features for 35,000 individuals, split into two groups by income. *Sampled Mixtures* (synthetic) contains 10,000 points from a Gaussian mixture with covariance τ in $\mathbb{R}^{15}$, simulating point clouds suited to k-means. The continuous distributions are *Gaussians* and *fragmented-hypercube* [Forrow et al., 2019],[2] with closed-form W_2; see Appendix 1 for details and more experiments.

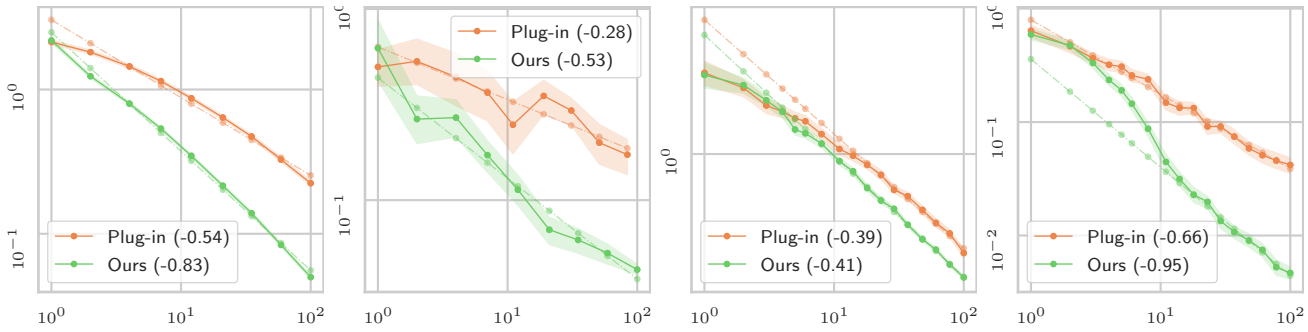
4.1 ALGORITHM 1

For each dataset, we compare the behavior of the plug-in estimator and that of Algorithm 1. We plot the mean *relative* error $\mathbb{E}_{X_n,\hat{S}_k}\left[|\text{Est}(k, k^2\log k) - W_2(\mu,\nu)|\right]/W_2(\mu,\nu)$, estimating the expectation with 100 runs. We display two types of plots: *(i)* mean relative error vs. k (size of the point clouds passed to the LP) (Figures 1, 2) and *(ii)* mean relative error vs. CPU time (Figure 3).

Results. Our estimator exhibits favorable behavior when estimating W_2 between large *point clouds*. In this case, the sample complexity of the plug-in estimator $W_2(\mu, \hat{\mu}_k)$ decays in $O(k^{-1/2})$, independently of the dimension or number of samples (these only affect the constant [Sommerfeld et al., 2018]), but ours enjoys a faster decay rate exponent—up to twice better. For continuous distributions, our results are similarly advantageous in the finite-sample regime for clusterable distributions but tend to the sample complexity rate in higher dimensions. They provide a way to verify Theorem 2 and to illustrate the different regimes. We notice in practice that oversampling enables the estimator to have much lower variance (fig. 7, supplement).

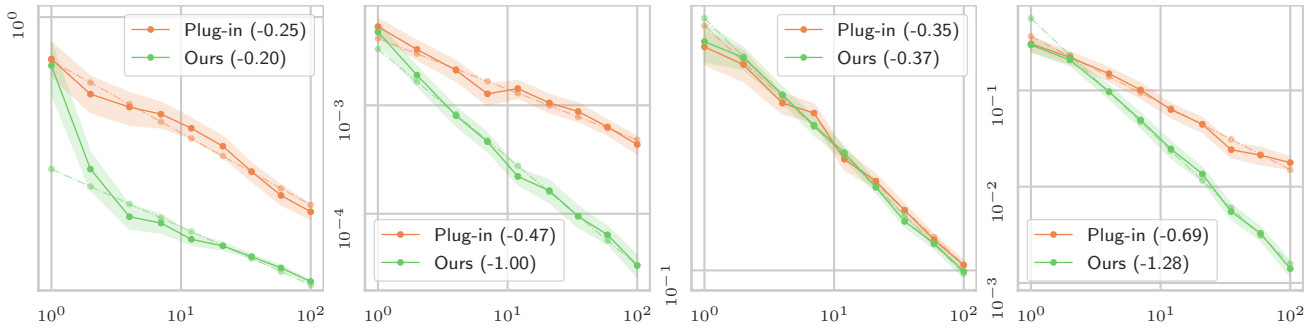
Discrete datasets. On the real-world datasets, the bias decays 45% (*DOT*, fig. 1a) to 65% (*Adult*, fig. 1b) faster. A simple analysis explains this: On a 100×100 image, with $k \le 100$ samples the plug-in estimator will sample $\sim 1\%$ of the image, whereas our estimator processes all the pixels and then subsamples the 100 most relevant. Synthetic experiments slightly qualify this analysis: When the data is well-clustered the improvement is up to twice the decay rate (fig. 1d), as expected from Proposition 2; however, when the point cloud is more spread out, the decay rate only marginally improves over plug-in estimation.

[2] What they refer to as "k-means & OT" is *not* our Algorithm 1, since they set $k = 4$. Their x-axis does not relate to overall computational complexity.



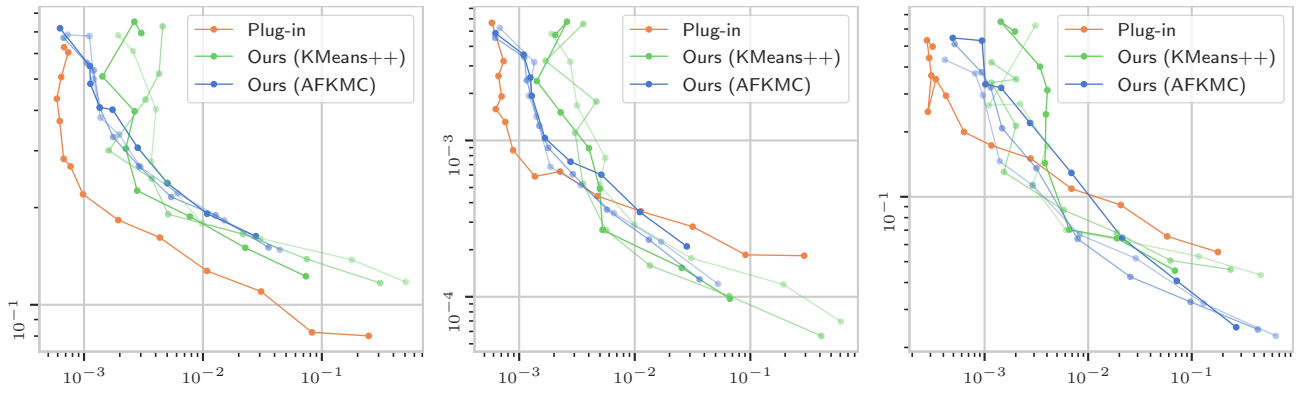
(a) DOT dataset. (b) Adult dataset. (c) Discrete mixture, large variance ($\tau = 0.1$). (d) Discrete mixture, low variance ($\tau = 10^{-4}$).

Figure 1: Mean relative error vs. k on discrete datasets. Values in parentheses display the regression coefficient computed for the second half of the graph. In *(a)*, we plot the average value of the 45 pairwise estimation on the DOT dataset ("Microscopy" images, 64 resolution).



(a) Gaussian with 10^{-1} diagonal covariance. (b) Gaussian with 10^{-4} diagonal covariance. (c) Fragmented hypercube, $d = 8$. (d) Fragmented hypercube, $d = 2$.

Figure 2: Mean relative error vs. k on continuous distributions. Values in parentheses display the regression coefficient computed for the second half of the graph. Left: Gaussian in $\mathbb{R}^5$. When the clusterable assumption does not hold, the improvement is negligible. However, when the finite sample rate is applicable, the improvement is striking ($\times 2.1$). Right: Fragmented hypercube [Forrow et al., 2019]. In high dimension, it resembles the uniform distribution and we get no improvement. In small dimension, the improvement is significant ($\times 1.8$).



(a) Spread Gaussian (b) Peaked Gaussian (c) Adult Dataset

Figure 3: Mean relative error vs. CPU time (s) on (a) Gaussians with unit covariance, (b) Gaussians with 10^{-4} diagonal covariance, (c) Adult dataset. One line corresponds to log-spaced values of k. Line's transparency correspond to various $\kappa \in \{1, 0.5, 0.1\}$, darkest for biggest value. Line's color corresponds to various **estimator**. We compare the plug-in estimator (orange) to two variants of our algorithm : k-means++ (green) or AFK-MC^2 (blue) from [Bachem et al., 2016] as a preprocessing step. For data with small quantization error, our approximate k-means pre-processing (even unoptimized) provides a clear advantage.

Continuous distributions. The plug-in estimator on Gaussian data recovers the expected $-1/d$ rate exponent when variance is high (fig. 2a); when the variance is low, we find the better finite sample complexity rate of $-1/2$ predicted by [Weed and Bach, 2019]. In this regime, our estimator beats the plug-in estimator by a large margin (fig. 2b). Asymptotically, both curves should reach the same slope of $-1/d$. Similarly, we should expect our estimator to degrade on the uniform distribution: for uniformly-spread data, quantization error decays in $k^{-1/d}$. The *Fragmented Hypercube* example confirms this: When $d = 2$, the distribution is clusterable (fig. 2d), but as d increases the quantization error is relatively high, eventually reaching the performance of the plug-in estimator (fig. 2c).

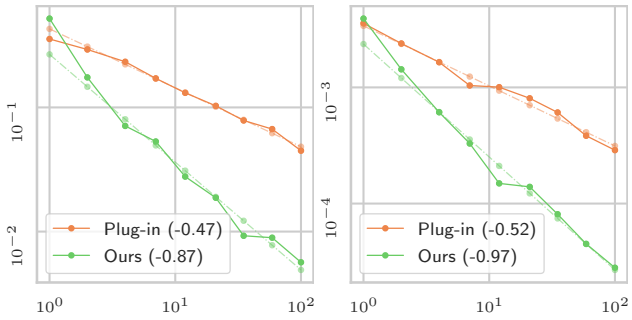
CPU time. Since our goal is to provide a faster W_2 approximation, we check the decay of the bias against CPU time. These experiments evaluate to what extent the theoretical improvement of the bias may be cancelled by overhead in k-means computation. The solver we use for OT [Flamary and Courty, 2017] is thoroughly optimized, making the comparison difficult. However, our estimator is only slower by a constant on spread out data (Figure 3(a)) and provides a clear advantage on clustered (Figure 3(b)) and real data (Figure 3(c)). To further improve, *(i)* our basic implementation of k-means++ could be optimized and *(ii)* we can use theoretically weaker minimizers of the quantization problem. In Figure 3, we use a faster approximate quantizer, AFK-MC^2 [Bachem et al., 2016] with fixed chain length on $n = k^2 \log k$ points (blue), which has overall complexity $k^2 \log k$ but weaker guarantees on the quantization error. Another alternative is to multiply the number of points used to compute the anchors (we tested $\kappa \in \{1, 0.5, 0.1\}$) to further decrease the complexity constant between the preprocessing and the OT estimation steps. This can be used as a hyper-parameter to balance faster execution with lower bias improvement. For these experiments, we use an Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz processor, with 8 GB memory. The k-means and OT solvers are implemented in C and wrapped in Python.

Variance of the estimator. Algorithm 1 relies on *oversampling*. Thus, we expect and confirm experimentally that it benefits from much lower variance compared to the plug-in estimator, as illustrated by the confidence intervals in Figures 1, 2 (plots are in log-log scale). For a more quantitative analysis, we plot the **empirical standard deviation** of Algorithm 1 on the Gaussian dataset on Figure 4. It is worth noticing that it exhibits a much lower variance no matter how clusterable the underlying distribution is. However, proving this requires bounding the stability of the optimal quantization solution, for which no directly applicable results exist.

Lloyd's algorithm. k-means++ is often used as an initialization step for Lloyd's algorithm. The latter converges to a local minimizer of the quantization error, at the expense of few more passes through the data, for an overall complexity of $O(nki)$, where i is the number of iterations. Theoretically, this algorithm makes the quantization error decay by $\log k$ at best. We verify experimentally that the improvement is marginal in Figure 5.

4.2 ALGORITHM 2

To test the performance of Algorithm 2, we compare it to do an approximate solver for entropy-regularized optimal transport, which is arguably the most popular occurence



(a) Gaussian with unit diagonal covariance (b) Gaussian with 10^{-4} diagonal covariance

Figure 4: Empirical standard deviation of Algorithm 1 vs. k. Sampling $k^2 \log k$ samples instead of k, our estimator manages a much lower standard deviation, independently of the clusterability of the distribution.

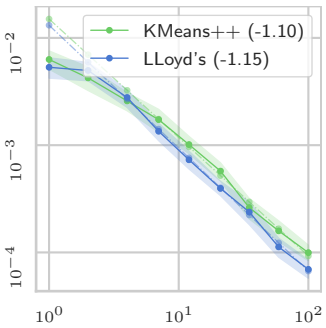


Figure 5: Mean relative error vs. k on the Gaussian dataset for Algorithm 1 (green), and the same algorithm succeeded by Lloyd's procedure (blue). The improvement of the latter is marginal and comes at the expense of few $O(nk)$ steps.

in machine learning applications. Specifically, for datasets (μ_n, ν_n), we measure the CPU time to execute Algorithm 2 with input $(\mu_n, \nu_n, \varepsilon)$ and APPROXOT$(\mu_n, \nu_n, 3\varepsilon)$, which are both guaranteed to output a $3\varepsilon-$approximation of OT. Here, APPROXOT is from [Altschuler et al., 2017], but any other approximate solver satisfying the same constraints on the input/output can be used. We display two types of plots: *(i)* CPU time vs. precision ε and *(ii)* estimated transport cost vs. precision ε. The former demonstrates efficiency while the latter shows that the output is indeed at most ε away from the unregularized cost.

Results. From the CPU time plots in fig. 6 (left column) the speedup introduced by our algorithm is unmistakable. It only matches the performance of APPROXOT for low values of ε, when QUANTIZE simply outputs the whole dataset to have a small enough quantization error. That's why it is most useful for structured data, e.g. peaked distributions (fig. 6.c) or real-world datasets (fig. 6.e) The error vs. ε plots (right column) suggest that the bounds in [Altschuler et al., 2017] are loose, since the error is often smaller than the guaranteed ε. Quantization enables us to have maximum efficiency for bounded inaccuracy.

5 CONCLUSION

Our algorithm is designed with practicality in mind: at best—and in most of our experiments—we observe and expect reduced bias for fixed computational budget; at worst, it behaves like plug-in estimation. Our bounds explain the estimator's good behavior by relating W_2 to quantization error. Even when we fall back to the $^{-1}/_{d}$ rate asymptotically, we have up to twice the decay rate in the finite sample case. Quantization is also efficient in approximate OT solvers, as it can match their error with improved time/space complexity.

Acknowledgements

This work was conducted in large part during an internship of Gaspard Beugnot at MIT.

The MIT Geometric Data Processing group acknowledges the generous support of Army Research Office grant W911NF2010168, of Air Force Office of Scientific Research award FA9550-19-1-031, of National Science Foundation grant IIS-1838071, from the CSAIL Systems that Learn program, from the MIT–IBM Watson AI Laboratory, from the Toyota–CSAIL Joint Research Center, from a gift from Adobe Systems, from an MIT.nano Immersion Lab/NCSOFT Gaming Program seed grant, and from the Skoltech–MIT Next Generation Program.

References

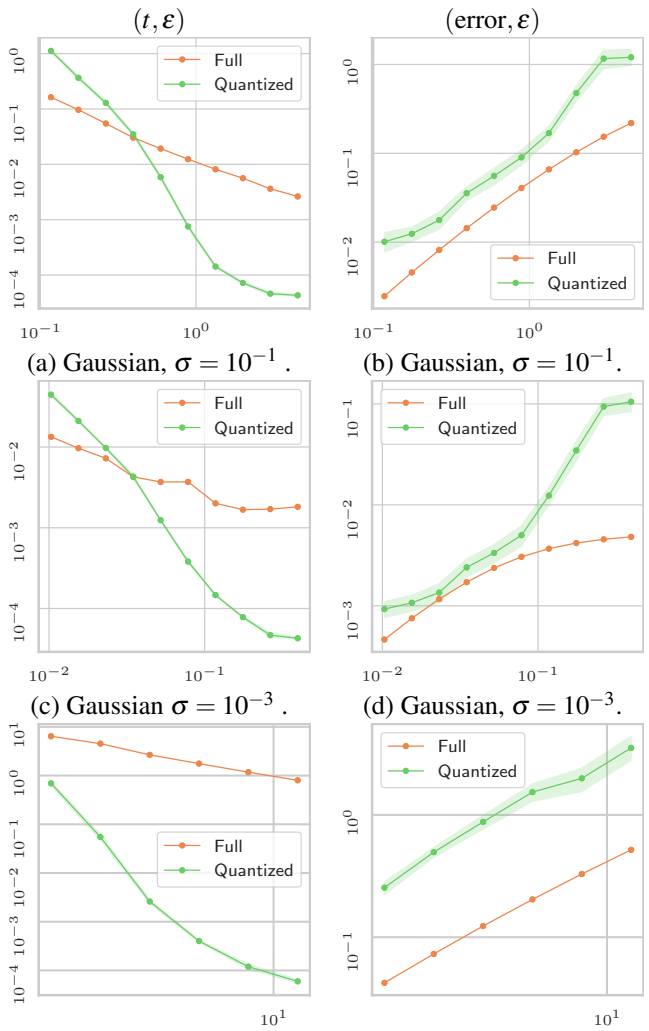
Jason Altschuler, Jonathan Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *CoRR*, abs/1705.09634, 2017. URL `http://arxiv.org/abs/1705.09634`.

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

Olivier Bachem, Mario Lucic, Hamed Hassani, and Andreas Krause. Fast and provably good seedings for k-means. In *Advances in Neural Information Processing Systems*, pages 55–63, 2016.

Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar, and Sergei Vassilvitskii. Scalable k-means++. *Proceedings of the VLDB Endowment*, 5(7), 2012.

Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 2019.



(a) Gaussian, $\sigma = 10^{-1}$. (b) Gaussian, $\sigma = 10^{-1}$.

(c) Gaussian $\sigma = 10^{-3}$. (d) Gaussian, $\sigma = 10^{-3}$.

(e) Adult Dataset. (f) Adult Dataset.

Figure 6: Left: CPU time (s) vs. ε, for Algorithm 2 and APPROXOT [Altschuler et al., 2017]. Right: absolute error vs. ε. The smallest precision for the range of ε is taken so that APPROXOT requires $n_{\max} = 10^4$ iterations. Algorithm 2 consistently provides an approximate solution an order of magnitude faster than APPROXOT.

Rainer Burkard, Mauro Dell'Amico, and Silvano Martello. *Assignment Problems, revised reprint*, volume 106. SIAM, 2012.

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

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

Richard Mansfield Dudley. The speed of mean Glivenko–Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

Pavel E. Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by sinkhorn's algorithm. *CoRR*, abs/1802.04367, 2018. URL `http://arxiv.org/abs/1802.04367`.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

Rémi Flamary and Nicolas Courty. POT: Python optimal transport library, 2017. URL `https://pythonot.github.io/`.

Aden Forrow, Jan-Christian Hütter, Mor Nitzan, Philippe Rigollet, Geoffrey Schiebinger, and Jonathan Weed. Statistical optimal transport via factored couplings. In *International Conference on Artificial Intelligence and Statistics*, pages 2454–2465, 2019.

Joel Franklin and Jens Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and Its Applications*, 114:717–735, 1989.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, 2018.

Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1574–1583, 2019.

Samuel Gerber and Mauro Maggioni. Multiscale strategies for computing optimal transport. *arXiv preprint arXiv:1708.02469*, 2017.

Ziv Goldfeld and Kristjan Greenewald. Gaussian-smooth optimal transport: Metric structure and statistical efficiency. *AISTATS*, 2020.

Benoit Kloeckner. Approximation by finitely supported measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 18(2):343–359, 2012.

Tianyi Lin, Nhat Ho, and Michael I. Jordan. On the efficiency of the Sinkhorn and Greenkhorn algorithms and their acceleration for optimal transport, 2019.

Stuart Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

David Pollard. Quantization and the method of k-means. *IEEE Transactions on Information Theory*, 28(2):199–205, 1982.

Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018.

Bernhard Schmitzer and Christoph Schnörr. A hierarchical approach to optimal transport. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 452–464. Springer, 2013.

Jorn Schrieber, Dominic Schuhmacher, and Carsten Gottschlich. DOTmark—A benchmark for discrete optimal transport. *IEEE Access*, 5:271–282, 2017. ISSN 2169-3536. doi: 10.1109/access.2016.2639065. URL `http://dx.doi.org/10.1109/ACCESS.2016.2639065`.

Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

Max Sommerfeld, Jörn Schrieber, Yoav Zemel, and Axel Munk. Optimal transport: Fast probabilistic approximation with exact solvers, 2018.

Cédric Villani. *Topics in Optimal Transportation*. Number 58. American Mathematical Society, 2003.

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