Dimensionality Reduction for Wasserstein Barycenter

Zachary Izzo
Stanford University
zle.izzo@gmail.com

Sandeep Silwal
MIT
silwal@mit.edu

Samson Zhou
Carnegie Mellon University
samsonzhou@gmail.com

Abstract

The Wasserstein barycenter is a geometric construct which captures the notion of centrality among probability distributions, and which has found many applications in machine learning. However, most algorithms for finding even an approximate barycenter suffer an exponential dependence on the dimension d of the underlying space of the distributions. In order to cope with this "curse of dimensionality," we study dimensionality reduction techniques for the Wasserstein barycenter problem. When the barycenter is restricted to support of size n, we show that randomized dimensionality reduction can be used to map the problem to a space of dimension $O(\log n)$ independent of both d and k, and that *any* solution found in the reduced dimension will have its cost preserved up to arbitrary small error in the original space. We provide matching upper and lower bounds on the size of the reduced dimension, showing that our methods are optimal up to constant factors. We also provide a coreset construction for the Wasserstein barycenter problem that significantly decreases the number of input distributions. The coresets can be used in conjunction with random projections and thus further improve computation time. Lastly, our experimental results validate the speedup provided by dimensionality reduction while maintaining solution quality.

1 Introduction

The Wasserstein barycenter (WB) is a popular method in statistics and machine learning for summarizing data from multiple sources while capturing their underlying geometry [AC11a]. The problem is defined as follows. Suppose we have a collection of data, represented as k discrete probability distributions $\mu_1, \ldots, \mu_k$ on $\mathbb{R}^d$. Given a set of non-negative weights $\lambda_1, \ldots, \lambda_k$ that sum to 1, and a class $\mathbb{P}$ of probability distributions on $\mathbb{R}^d$, a Wasserstein barycenter under the L_p objective for a parameter $p > 0$ is a probability distribution $\nu \in \mathbb{P}$ that minimizes

$$\sum_{i=1}^{k} \lambda_i W_p(\mu_i, \nu)^p, \tag{1}$$

where $W_p(\mu_i, \nu)$ is the p-Wasserstein distance.

The Wasserstein barycenter is a natural quantity that captures the geometric notion of centrality among point clouds, as it utilizes the optimal transport distance [BT97] between a number of observed sets. Thus, Wasserstein barycenters have been extensively used in machine learning [SLD18], data sciences [RU02, EHJK20], image processing [RGT97], computer graphics [PW09], and statistics [Vil08], with applications in constrained clustering [CD14, HNY+17], Bayesian learning [SLD18], texture mixing [RPDB11], and shape interpolation [SDGP+15].

Unfortunately, the problem is NP-hard to compute [AB21, BP21] and many algorithms that even approximate the Wasserstein barycenter suffer from large running times, especially if the datasets are high dimensional [MC19]. Indeed, [ABA21] recently gave an algorithm that computes the

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Wasserstein barycenter using runtime that depends exponentially on the dimension, thus suffering the "curse of dimensionality."

To alleviate these computational constraints, we consider dimensionality reduction for computing the Wasserstein barycenter. Dimensionality reduction can be used to improve the performance of downstream algorithms on high dimensional datasets in many settings of interest, e.g., see the survey [CG15]. In the specific case of Wasserstein barycenters, dimensionality reduction has several practical and theoretical benefits, including lower storage space, faster running time in computing distances, and versatility: it can be used as a pre-processing tool and combined with *any* algorithm for computing the Wasserstein barycenter.

1.1 Our Results

In this paper, we study dimensionality reduction techniques for computing a Wasserstein barycenter of discrete probability distributions. Our main results show that it is possible to project the distributions into low dimensions while provably preserving the quality of the barycenter. A key result in dimensionality reduction is the classical Johnson-Lindenstrauss (JL) lemma [JL84], which states that projecting a dataset of N points into roughly $O(\log N)$ dimensions is enough to preserve all pairwise distances.

Using the JL lemma, we first show that we can assume the distributions lie in $O(\log(nk))$ dimensions, where k is the number of input distributions whose barycenter we are computing, n is the size of the support of the barycenter, and each of the k input distributions has support size $\mathrm{poly}(n)$. For $p = 2$, there exists a closed form for the cost of any candidate barycenter in terms of the pairwise distances of the points in the input distributions. Thus it is straightforward to see that our bound results from the fact that there are $k \cdot \mathrm{poly}(n)$ total points masses in the union of all the distributions and therefore, projecting them into a dimension of size $O(\log(k\,\mathrm{poly}(n))) = O(\log(nk))$ suffices to preserve all of their pairwise distances. However for $p \neq 2$, a closed form for the optimal cost no longer exists, so preservation of all pairwise distances is insufficient. Instead, we make use of a Lipschitz extension theorem, namely the Kirszbraun theorem, which allows us to "invert" the dimensionality reduction map and argue the preservation of the cost of the Wasserstein barycenter under a general L_p objective. For more details, see Section 3.

Dimensionality reduction independent of k. While the JL lemma is known to be tight [LN16, LN17], it is possible to improve its dimensionality guarantees for *specific* problems, such as various formulations of clustering [CEM+15, BBC+19, MMR19]. Indeed, our main result is that we can achieve a dimension bound *beyond* the $O(\log(nk))$ bound that follows from the JL lemma and Kirszbraun theorem. We show that it suffices to project the support points onto $O(\log n)$ dimensions, which is *independent* of the number of distributions k. In fact, we show a stronger statement that projecting the points supported by the distributions onto $O(\log n)$ dimension preserves the cost of the objective (1) for *any* distribution ν supported on at most n points (Theorem 4.1). The algorithmic application of this theorem is that one can take *any* approximation algorithm or heuristic for computing the Wasserstein barycenter and combine it with dimensionality reduction. A simplification of our theorem is stated below where we omit some parameters for clarity.

Theorem 1.1 (Theorem 4.1 Simplified)**.** *Let $\mu_1, \ldots, \mu_k$ be discrete probability distributions on $\mathbb{R}^d$ such that $|supp(\mu_i)| \leq \mathrm{poly}(n)$ for all i. There exists a dimensionality reduction map $\pi : \mathbb{R}^d \to \mathbb{R}^m$ for $m = O(\log n)$ such that projection under π preserves the cost of objective* (1) *for any ν supported on at most n points.*

The result is surprising because the projected dimension is independent of the number of input distributions k, which could be significantly larger than n. Thus the random projection map π can no longer even guarantee the preservation of a significant fraction of pairwise distances between the support points of the k distributions. Our main tool is a "robust" Lipschitz extension theorem introduced in [MMR19] for k-means clustering. We adapt this analysis to the geometry of the Wasserstein barycenter problem.

Optimality of dimensionality reduction. We complement our upper bound results by showing that our dimension bound of $\log n$ dimensions is *tight* if a random Gaussian matrix is used as the projection map. We also show that the JL lemma is tight for the related problem of computing the optimal transport between two distributions with support of size n. More specifically, we give a lower bound showing that $\Omega(\log n)$ dimension is needed for a random projection to preserve the optimal

transport cost. Thus our results show a separation between the geometry of the optimal transport problem and the geometry of the Wasserstein barycenter problem, as we overcome the JL bound in the latter.

Hardness of approximation. In addition, we also show the NP-hardness of approximation for the Wasserstein barycenter problem. Namely, we show that it is NP-hard to find an approximate barycenter that induces a cost that is within a factor of 1.0013 of the optimal barycenter if we restrict the support size of the barycenter. This complements recent work of [AB21, BP21], who showed that computing sparse Wasserstein barycenters is NP-hard.

Coresets for Wasserstein barycenters. An alternate way to reduce the complexity of datasets is through the use of coresets, which decrease the effective data size by reducing the number of input points rather than the input dimension d. If the number of input distributions k is significantly larger than the support size n, we show that there exists a weighted subset C of roughly $\mathrm{poly}(n)$ distributions, so that computing the optimal barycenter on C is equivalent to computing the optimal barycenter on the original input up to a small approximation loss. Hence, it can potentially be much more efficient to use the subset C in downstream algorithms involving Wasserstein barycenters. Moreover, the coreset is not mutually exclusive with our techniques for reducing the ambient dimension d. Our techniques show that we can simultaneously reduce both the size of the input distribution k and the dimension d of the data, while preserving the optimal clustering within a small approximation factor.

In Supplementary Section E, we also show a connection between the Wasserstein barycenter problem and constrained low-rank problems. This class of problems includes examples such as the singular value decomposition (SVD) and k-means clustering. While this connection does not yield any improved results, it classifies the Wasserstein barycenter as a member of a general class of problems, and this classification could have further applications in the future.

Experiments. Finally, we present experimental evaluation of our proposed methodology. Note that our results imply that we can use dimensionality reduction in conjunction with *any* Wasserstein barycenter algorithm and still roughly retain the approximation guarantees of the algorithm used. Specifically, we give examples of real high dimensional datasets such that solving the Wasserstein barycenter problem in a reduced dimension leads to computational savings while preserving the quality of the solution. Our experiments in Section 7 demonstrate that on natural datasets, we can reduce the dimension by 1-2 orders of magnitude while increasing the solution cost by only 5%. We also empirically test our coreset construction. Our method both reduces error and requires fewer samples than simple uniform sampling.

1.2 Related Work

[AB21, BP21] showed that computing sparse Wasserstein barycenters is NP-hard; hence, most of the algorithmic techniques focus on computing approximate Wasserstein barycenters that induce a cost within an additive ε of the optimal cost. [AC11b] first considered approximating Wasserstein barycenters when either (1) the distributions $\mathbb{P}$ only have discrete support on $\mathbb{R}$, (2) $k = 2$, or (3) the distributions μ_i are all multivariate Gaussians in $\mathbb{R}^d$. Although there is a line of research that studies the computation of barycenters of continuous distributions, e.g. [ÁDCM16, CMRS20], we focus on discrete input distributions. For discrete input distributions, the majority of the literature can be categorized by its assumptions of the support of the barycenter [ABA21].

Fixed-support. The "fixed-support approximation" class of algorithms assume that the support of the barycenter is among a fixed set of possible points. It then remains for the algorithms to solve a polynomial-size linear program associated with the corresponding set [CD14, BCC+15, COO15, SCSJ17, KTD+19, LHC+20]. Unfortunately, the set of possible points must often be an ε-net over the entire space, which results in a size proportional to $1/\varepsilon^d$ that suffers from the curse of dimensionality. Nevertheless for constant dimension, the algorithms typically have runtime $\mathrm{poly}(n, k, D/\varepsilon)$, where D is an upper bound on the diameter of the supports of the input distributions. This is further improved by an algorithm of [ABA21] that achieves runtime $\mathrm{poly}(n, k, \log(D/\varepsilon))$.

Free support. A separate class of algorithms do not make assumptions about the possible support of the optimal barycenter. These "free-support algorithms" instead optimize over the entire set of up candidate barycenters, which can be as large as n^k in quantity. Thus these algorithms, e.g., [CD14, LSPC19], either use exponential runtime or a number heuristics that lack theoretical guarantees. [ABA21] showed how to explore the n^k possible points in polynomial time for fixed d.

2 Preliminaries

Notation. For a positive integer n, we denote $[n] := \{1, 2, \ldots, n\}$. We use $\mu_1, \ldots, \mu_k$ to denote the k distributions whose Wasserstein barycenter we wish to compute. While the Wasserstein barycenter problem is well defined for continuous distributions, in practice and in actual computations, the distributions μ_i are assumed to be discrete distributions that are supported on some number of point masses. This is also the assumption we make. More specifically, we assume that each of the distributions μ_i are discrete distributions supported on at most $T \leq n^C$ points where C is a fixed constant. That is, $\mu_i = \sum_{j=1}^{T} a(x_{ij})\delta_{x_{ij}}$, where δ_x is a delta function at x and $a(x)$ is the weight assigned to a point x in its corresponding μ_i. We note that if there is some point x in the support of more than one of the μ_is, then the weight function a may not be well-defined. Instead, we implicitly assume that $a = a(x, i)$ is a function of both the point *and* the distribution from which it comes, but we suppress this dependence on i for notational clarity.

The distribution ν denotes a candidate for the Wasserstein barycenter of the μ_i. We write $\nu = \sum_{j=1}^{n} b_j \delta_{\nu^j}$. In general, an actual Wasserstein barycenter (in the sense of minimizing the objective (1) over all possible ν of any support size) may have support size up to $|\bigcup_{i=1}^{k} \text{supp}(\mu_i)|$ [ABM16]. Throughout this paper, we will restrict ourselves to computing (approximate) barycenters of support size at most n. When we refer to an optimal barycenter, we mean a distribution that minimizes the objective (1) *within this restricted class*.

Problem description. The goal is to compute a distribution $\nu \in \mathbb{R}^d$, consisting of at most n point masses, to minimize the objective (1). As previously mentioned, $W_p(\mu_i, \nu)$ is the Wasserstein p-metric, defined as

$$W_p(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p d\gamma(x, y) \right)^{1/p}$$

where $\Gamma(\mu, \nu)$ is the set of all joint distributions with marginals μ and ν (i.e. all couplings of μ and ν) and $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^d$. When μ and ν are discrete distributions, $W_p(\mu, \nu)^p$ p-metric is the cost of the minimum cost flow from μ to ν with edge costs being the Euclidean distance raised to the p-th power. For simplicity, we assume that the distributions $\mu_1, \cdots, \mu_k$ are weighted equally (each $\lambda_i = 1/k$ in (1)) but our results hold in the general case as well. The most common choice of p is $p = 2$.

Description of ν. The barycenter ν can be characterized as follows. Recall that ν is supported on the points $\nu^1, \ldots, \nu^n$. For the optimal coupling of each μ_i to ν, let $w_j(x)$ denote the total weight sent from x (in the support of one of the μ_is) to ν^j. (The same note about suppressing the dependence of w_j on the distribution μ_i from which x comes applies here.) Let $S_j = \{x \in \bigcup_{i=1}^{k} \text{supp}(\mu_i) : w_j(x) > 0\}$ denote the set of all points in the μ_is with some weight sent to ν^j. Then *given* the set S_j and weighting function $w_j(\cdot)$, we can reconstruct ν^j since it must minimize the objective

$$\sum_{x \in S_j} w_j(x) \|x - \nu^j\|^p. \tag{2}$$

Indeed if ν^j does not minimize this quantity, we can change it and reduce the cost of (1).

Consider the case of $p = 2$. For a fixed j, (2) is just a weighted k-means problem whose solution is the weighted average of the points in S_j. To prove this, consider taking the gradient of (2) with respect to the k-th coordinate of ν^j. Then setting it equal to 0 gives us that the k-th coordinate will be the weighted average of the k-th coordinates of the points S_j. That is, we have

$$\nu^j = \frac{\sum_{x \in S_j} w_j(x) x}{\sum_{x \in S_j} w_j(x)} = \frac{1}{k b_j} \sum_{x \in S_j} w_j(x) x. \tag{3}$$

The second equality results from observing that in order for the w_js to define a proper coupling, we have $\sum_{j=1}^{n} w_j(x) = a(x)$ for all x in the support of the μ_is, and $\sum_{x \in \text{supp}(\mu_i)} w_j(x) = b_j$ for all i, along with $w_j(x) \geq 0$. In particular, this implies that $\sum_{x \in S_j} w_j(x) = k b_j$ for all $j = 1, \ldots, n$.

For arbitrary p, such a concise description of ν^j is not possible. Therefore an alternate, but equivalent, way to characterize the distribution ν is to just define the sets S_j and weight functions $w_j(\cdot)$ for $1 \leq j \leq n$. This motivates the following definitions.

Definition 2.1. *A solution* $(S, w) = (S_1, \ldots, S_n, w_1, \ldots, w_j)$ *is a valid partition as described previously (meaning that these partitions come from the optimal coupling between each* μ_i *to a* fixed ν*), along with the corresponding weight functions* $w_j(\cdot)$.

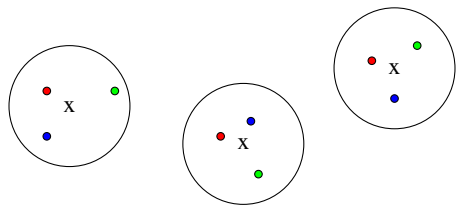


Figure 1: Points of the same color belong to the same distribution. The sets S_j are denoted by the large black circles. Given the partitions S_j (denoted by large black circles) and associated weight functions w_j, we can reconstruct the barycenter (denoted by crosses).

Definition 2.2. *Let* (S, w) *be a solution. The cost of this solution, denoted* $\text{cost}_p(S)$*, is the value of the objective* (1) *when we reconstruct* ν *from* S *and* w *and evaluate* (1)*:*

$$\text{cost}_p(S) = \min_{\nu} \frac{1}{k} \sum_{j=1}^{n} \sum_{x \in S_j} w_j(x) \|x - \nu^j\|^p.$$

Similarly for a projection π*,* $\text{cost}_p(\pi S)$ *denotes the value of the objective* (1) *when we first project each of the distributions to* $\mathbb{R}^m$ *using* π*, then compute* $\widetilde{\nu}$ *using the original weights* w_j*:*

$$\text{cost}_p(\pi S) = \min_{\widetilde{\nu}} \frac{1}{k} \sum_{j=1}^{n} \sum_{x \in S_j} w_j(x) \|\pi(x) - \widetilde{\nu}^j\|^p.$$

Note that each $\widetilde{\nu}^j \in \mathbb{R}^m$*. We suppress the dependence of the cost on* w *for notational convenience.*

For the case of $p = 2$, we can further massage the value of ν^j in (2). Let $\bar{x}$ denote the weighted average of points in S_j (given by (3)). From our discussion above, we know that $\nu^j = \bar{x}$. After some standard algebraic manipulation, we can show that $\sum_{x \in S_j} w_j(x) \|x - \nu^j\|^2 = \sum_{x \in S_j} w_j(x) \|x\|^2 - k b_j \|\bar{x}\|^2$ and $\sum_{x,y \in S_j} w_j(x) w_j(y) \|x - y\|^2 = 2k b_j \left(\sum_{x \in S_j} w_j(x) \|x\|^2 - k b_j \|\bar{x}\|^2 \right)$. Combining these equations yields the following for the $p = 2$ objective.

$$\frac{1}{2k b_j} \left(\sum_{x,y \in S_j} w_j(x) w_j(y) \|x - y\|^2 \right) = \sum_{x \in S_j} w_j(x) \|x - \nu^j\|^2. \tag{4}$$

Dimension reduction. In this paper we are concerned with dimensionality reduction maps $\pi : \mathbb{R}^d \to \mathbb{R}^m$ that are JL projections, i.e., any dimensionality reduction map that satisfies the condition of the JL lemma. This includes random Gaussian and sub-Gaussian matrices [LN16, MMR19]. We are mainly concerned with making the projection dimension m as small as possible.

Consider any algorithm $\mathcal{A}$ that, given $\mu_1, \cdots, \mu_k$, solves for some approximate or exact ν minimizing the objective (1). We can combine any such $\mathcal{A}$ with dimensionality reduction by first projecting the point masses of the μ_i down to $\mathbb{R}^m$ for some $m < d$ and using $\mathcal{A}$ to compute some barycenter $\widetilde{\nu}$ in $\mathbb{R}^m$. Then, we can consider the solution (S, w) induced by $\widetilde{\nu}$ (see Definitions 2.1 and 2.2) to *reconstruct* the appropriate ν in the original dimension $\mathbb{R}^d$ using the objective Eq. (2). Note that this objective is a convex program for any $p \geq 1$ since we are given S_j and $w_j(\cdot)$. For $p = 2$ (which is the most common case), ν^j has a particularly simple form which is the weighted average of the points in S_j (see Eq. (3)). This procedure is outlined in Algorithm 1.

As a corollary of our results, if algorithm $\mathcal{A}$ takes time $T(n, k, d)$, then using dimensionality reduction as in the procedure outlined above takes time $T(n, k, m)$ plus the time to perform the projection and reconstruct the barycenter using the solution S. The cost of running algorithm $\mathcal{A}$ is usually much more expensive than performing the projection, and the reconstruction step can also be solved efficiently since it is convex. In the case of $p = 2$, the reconstruction just amounts to computing n weighted means. Therefore for $m \ll d$, we get significant savings since $T(n, k, m) \ll T(n, k, d)$.

Algorithm 1 Using dimensionality reduction with any algorithm $\mathcal{A}$ for computing WB

Require: k discrete distributions $\mu_1, \cdots, \mu_k$ with point masses in dimension $\mathbb{R}^d$, projection dimension m, algorithm $\mathcal{A}$

Project the point masses of each distribution μ_i to dimension $\mathbb{R}^m$ using a JL projection
Use algorithm $\mathcal{A}$ to solve (or approximately solve) the Wasserstein barycenter problem in $\mathbb{R}^m$ to get a distribution $\widetilde{\nu}$ `//`$\widetilde{\nu}$ `is a discrete distribution in` $\mathbb{R}^m$
Let (S, w) be the solution that partitions the the point masses of the distributions as described in Definition 2.1
for each $S_j \in S$ **do**
  Solve for ν^j minimizing $\sum_{x \in S_j} w_j(x) \|x - \nu^j\|^p$ `//This is a convex program for` $p \geq 1$. `For` $p = 2$, ν^j `is just the weighted average of points in` S_j.
end for
Output the distribution ν supported on ν^j, and where ν^j has the same weight as $\widetilde{\nu}^j$

3 Reduction to $O(\log(nk))$ Dimensions

We first show that it suffices to project the point masses of the input distribution into $O(\log(nk))$ dimensions and guarantee that the cost of any *solution* is preserved. Note that our results hold *simultaneously for all solutions.* We first state the $p = 2$ case.

Theorem 3.1. *Consider a JL projection π from $\mathbb{R}^d$ to $\mathbb{R}^m$ for $m = O(\log(nk/\delta)/\varepsilon^2)$. Then*

$$\mathbb{P}\left(\text{cost}_2(\pi S) \in \left[(1-\varepsilon)^2 \cdot \text{cost}_2(S), (1+\varepsilon)^2 \cdot \text{cost}_2(S)\right] \textit{ for all solutions } S\right) \geq 1 - \delta.$$

Proof. The proof follows from the solution decomposition given in (4) if we condition on all the pairwise distances being preserved which happens with probability $1 - \delta$. □

A decomposition similar to (4) does not exist for $p \neq 2$. To prove an analogous theorem for $p \neq 2$, we need the following Lipschitz extension theorem which roughly allows us to "invert" a dimensionality reduction map.

Theorem 3.2 (Kirszbraun Theorem [Kir34])**.** *For any $D \subset \mathbb{R}^m$, let $f : D \to \mathbb{R}^d$ be an L-Lipschitz function. Then there exists some extension $\widetilde{f} : \mathbb{R}^m \to \mathbb{R}^d$ of f to the entirety of $\mathbb{R}^m$ such that $f(x) = \widetilde{f}(x)$ for all $x \in D$ and $\widetilde{f}$ is also L-Lipschitz.*

The Kirszbraun theorem allows us to prove Theorem 3.1 for general p with a dimension bound of $m = O(\log(nk/\delta)p^2/\varepsilon^2)$ (see Theorem A.1 in Supplementary Section A).

The overview for the proof strategy for the general $p \neq 2$ case is as follows. First suppose that all the pairwise distances between the support points of all the distributions are preserved under the projection map up to multiplicative error $1 \pm \varepsilon$. This event happens with probability at least $1 - \delta$. We then consider the map $f : \mathbb{R}^m \to \mathbb{R}^d$ that maps each of the *projected* points to its original counterpart in $\mathbb{R}^d$. Note that the map is from the *smaller* dimension m to the larger dimension d. On the support points, we know that f is $(1 + \varepsilon)$-Lipschitz by our assumption above.

Now if the projection caused the cost of πS to decrease significantly, then using the Kriszbraun theorem, one could "lift" the corresponding barycenter $\widetilde{\nu}$ from the projected dimension to the original dimension using the extension map $\widetilde{f}$. Then since $\widetilde{f}$ is Lipschitz, this lifted barycenter $\widetilde{f}(\widetilde{\nu})$ plugged into Eq. (2) would subsequently have cost smaller than the original barycenter that corresponds S in the original dimension. This is a contradiction in light of Eq. (2) and the description of ν given in Section 2. Note that the exact description of $\widetilde{f}$ does not matter for the analysis, just that such a map exists. A complete, rigorous proof can be found in the supplementary section.

4 Optimal Dimensionality Reduction

We now present our main theorem which improves the guarantees of Theorems 3.1 and A.1.

Theorem 4.1. *Let $\mu_1, \ldots, \mu_k$ be discrete probability distributions on $\mathbb{R}^d$ such that $|supp(\mu_i)| \leq \mathrm{poly}(n)$ for all i. Let $d \geq 1$, $\varepsilon, \delta \in (0,1)$, and $p \geq 1$. Let $\pi_{d,m} : \mathbb{R}^d \to \mathbb{R}^m$ be a family of random JL maps with $m = O\left(\frac{p^4}{\varepsilon^2} \log \frac{n}{\varepsilon\delta}\right)$. Then we have,*

$$\mathbb{P}\left(\mathrm{cost}_p(\pi S) \in \left[(1-\varepsilon)\cdot \mathrm{cost}_p(S), (1+\varepsilon)\cdot \mathrm{cost}_p(S)\right] \textit{ for all solutions } S\right) \geq 1-\delta.$$

We now give an overview of the proof strategy for Theorem 4.1, deferring all technical details to the supplementary section. Ideally, one would like to use a strategy similar to the proof of Theorem A.1. The key bottleneck is that when we project down to the m specified in Theorem 4.1, a large number of pairwise distances between the support points of the k distributions can be distorted (since we are projecting to a dimension smaller than $O(\log(nk))$). Therefore, the Kirszbraun theorem cannot apply as the map f described in the proof strategy of Theorem A.1 is no longer Lipschitz on the support points.

To overcome this barrier, we generalize an approach of [MMR19], who achieved the optimal dimensionality bounds for k-means clustering beyond the naïve JL bound by defining a *distortion graph* on the set of input points, which has an edge between each pair of points if their pairwise distance is distorted by at least a $(1+\varepsilon)$-factor under the random projection map π. They show that the distortion graph is everywhere sparse, i.e., each vertex has small expected degree in the distortion graph, which implies a "robust" Kirszbraun theorem (for their particular problem of k-means clustering). Namely, there exists an extension map $\widetilde{f} : \mathbb{R}^d \to \mathbb{R}^m$ and a specific point $v \in \mathbb{R}^m$ in the projected space such that a large fraction of the distances from the pre-image $\widetilde{f}^{-1}(v)$ to the input points in $\mathbb{R}^d$ are preserved. Moreover, the input points whose distance to $\widetilde{f}^{-1}(v)$ is not preserved can be shown to contribute small error to the k-means clustering cost.

The dimensionality reduction maps of Theorem 4.1 generally require multiplication by a dense matrix of (scaled) subgaussian variables. In the Supplementary Section, we show that "faster" dimensionality reduction maps can also be used by providing a trade off between the projection runtime and the dimension m. Note that in practice, performing the projection is extremely cheap since we only need to perform one matrix multiplication, which is highly optimized. Therefore the cost of any algorithm for Wasserstein barycenter will typically outweigh the cost of computing the projection.

4.1 Dimensionality Reduction Lower Bounds

In this section, we state lower bounds on the projection dimension m for the Wasserstein barycenter problem. Theorem 4.2 shows that Theorem 4.1 is tight up to constant factors.

Theorem 4.2. *Consider the setup of Theorem 4.1. Any Gaussian matrix used as a dimension reduction map that allows a $(1+\varepsilon)$-approximation to the optimal Wasserstein barycenter requires dimension $\Omega(\log n/\varepsilon^2)$.*

We also prove that one cannot do better than the naïve JL bound for the related problem of computing the optimal transport between two discrete distributions with n point masses each. This is in contrast to the case of Wasserstein barycenter where we were able to overcome the bound that comes from the JL lemma alone. Theorem 4.3 shows that the optimal solution in the projected dimension can induce a poor quality solution in the original dimension if the projection dimension is smaller than $\log n$.

Theorem 4.3. *There exists point sets $A, B \subset \mathbb{R}^d$ with $|A| = |B| = n$ and matching cost M between them, such that if randomly projected down to $m = o(\log n)$ dimensions using an appropriately scaled Gaussian random matrix, the pull back cost of the optimal matching in $\mathbb{R}^m$ is at least $\omega(M)$.*

In addition, we prove a related theorem which states that the *cost* of the optimal transport is heavily distorted if we project to fewer than $\log n$ dimensions.

Theorem 4.4. *There exists point sets $A, B \subset \mathbb{R}^d$ with $|A| = |B| = n$ and matching cost M between them, such that if randomly projected down to $m = o(\log n)$ dimensions using an appropriately scaled Gaussian random matrix, the cost of optimal matching in $\mathbb{R}^m$ is $o(M)$ with probability at least $2/3$.*

See Supplementary Section D for full proofs.

5 Coresets

In this section, we give a coreset construction for Wasserstein barycenters. Our goal is to reduce the number of distributions k to only depend polynomially on n. We first define our notion of coresets.

Definition 5.1 (Coreset)**.** *Fix $p \geq 1$. Let C and M be two sets of distributions in $\mathbb{R}^d$ where all distributions consist of $\mathrm{poly}(n)$ point masses. C is called an ε-*corset *for the set of distributions M if there exist weights w_c for $c \in C$ such that for all distributions ν of support size at most n, it holds that*

$$(1-\varepsilon)\sum_{c\in C} w_c\, W(c,\nu)^p \leq \frac{1}{|M|}\sum_{\mu\in M} W(\mu,\nu)^p \leq (1+\varepsilon)\sum_{c\in C} w_c\, W(c,\nu)^p.$$

The main result of this section is the following theorem.

Theorem 5.2 (Theorem C.6 simplified)**.** *Let M be a set of discrete distributions in $\mathbb{R}^d$, each supported on at most $\mathrm{poly}(n)$ point masses. There exists a weighted subset $K \subseteq M$ of size $\mathrm{poly}(n,d)/\varepsilon^2$ that satisfies Definition 5.1 for $p = O(1)$.*

To prove Theorem 5.2, we follow the "importance sampling" by sensitivities framework in conjunction with using structural properties of the Wasserstein barycenter problem itself. The sensitivity sampling framework has been successfully applied to achieve corsets for many problems in machine learning (see the references in the survey [BLK17]). Note that we have not attempted to optimize the constants in our proofs and instead focus on showing that k can be reduced to $\mathrm{poly}(n,d)$ for simplicity. The formal proof of Theorem 5.2 is deferred to the supplementary section.

We now describe the high level overview of the proof. We form the set C by sampling distributions in M with replacement based on their "importance" or contribution to the total cost. The notion of importance is formally captured by the definition of sensitivity.

Definition 5.3 (Sensitivity)**.** *Consider the set N of all possible barycenter distributions ν with support size at most n. The sensitivity of a distribution $\mu \in M$ is defined as*

$$\sigma(\mu) = \sup_{\nu\in N} \frac{W(\mu,\nu)^p}{\frac{1}{|M|}\sum_{\mu'\in M} W(\mu',\nu)^p}.$$

The total sensitivity is defined as $\mathfrak{S} = \frac{1}{|M|}\sum_{\mu\in M}\sigma(\mu)$.

To see why such a notion is beneficial, consider the case that one distribution μ consists of point masses that are outliers among all of the point masses comprising the distributions in M. Then it is clear that we must sample μ with a higher probability if we wish to satisfy the definition of a coreset. In particular, we sample each distribution in M with probability proportional to (an upper bound on) its sensitivity. Using a standard result in coreset construction, we can bound the size of the coreset in terms of the total sensitivity and a measure of the "complexity" of the Wasserstein barycenter problem which is related to the VC dimension. In particular, we utilize the notion of psuedo-dimension.

Definition 5.4 (Pseudo-Dimension, Definition 9 [LFKF18])**.** *Let $\mathcal{X}$ be a ground set and $\mathcal{F}$ be a set of functions from $\mathcal{X}$ to the interval $[0,1]$. Fix a set $S = \{x_1, \cdots, x_n\} \subset \mathcal{X}$, a set of reals numbers $R = \{r_1, \cdots, r_n\}$ with $r_i \in [0,1]$ and a function $f \in \mathcal{F}$. The set $S_f = \{x_i \in S \mid f(x_i) \geq r_i\}$ is called the induced subset of S formed by f and R. The set S with associated values R is shattered by $\mathcal{F}$ if $|\{S_f \mid f \in \mathcal{F}\}| = 2^n$. The* pseudo-dimension *of $\mathcal{F}$ is the cardinality of the largest shattered subset of $\mathcal{X}$ (or ∞).*

The following theorem provides a formal connection between the size of coresets and the notion of sensitivity and psuedo-dimension. Note that the statement of the theorem is more general and applies to a wider class of problems. However, we specialize the theorem statement to the case of Wasserstein Barycenters.

Theorem 5.5 (Coreset Size, Theorem 2.4.6 in [Lan18], Theorem 2.3 in [BLK17] for the case of Wasserstein Barycenters)**.** *Let $\varepsilon > 0$ and $\delta \in (0,1)$. Let $s : M \to \mathbb{R}^{\geq 0}$ denote any upper bound function on the sensitivity $\sigma(\cdot)$ defined in Definition 5.3 and let $S = \frac{1}{|M|}\sum_{\mu\in M} s(\mu)$. Consider a set K of $|K|$ samples of M with replacement where each distribution $\mu \in M$ is sampled with probability $q(\mu) = s(\mu)/(|M|\cdot S)$ and each sampled point is assigned the weight $1/(|M|\cdot|K|\cdot q(\mu))$. Let $\mathcal{F}$*

denote the set of functions

$$\mathcal{F} = \left\{ \frac{W(\cdot, \nu)^p}{Sq(\cdot) \sum_{\mu \in M} W(\mu, \nu)^p} \mid \nu \in N \right\}$$

where N is the set of all possible barycenter distributions with support size at most n. Let d' denote the pseudo-dimension of $\mathcal{F}$. Then the set K (along with the associated weights) satisfies Definition 5.1 with probability at least $1 - \delta$ if

$$|K| \geq \frac{cS}{\varepsilon^2} \left(d' \log S + \log \frac{1}{\delta} \right)$$

where $c > 0$ is some absolute constant.

Thus, the bulk of our work lies in bounding the sensitivities and psuedo-dimension. For the former quantity, we exploit the fact that the Wasserstein distance is a metric. The latter requires us to use tools from statistical learning theory which relate the VC dimension of a function class to its algorithmic complexity (see Lemmas C.3 and C.4). Full details given in Supplementary section C.

6 Other Theoretical Results

We now present some additional theoretical results pertaining to Wasserstein barycenters. Our first result is that Wasserstein barycenters can be formulated as a constrained low-rank approximation problem. This class of problems includes coputing the SVD and k-means clustering [CEM+15]. Formally, we prove the following theorem.

Theorem 6.1. *Given discrete distributions $\mu_1, \ldots, \mu_k \in \mathbb{R}^d$ with support size at most n, consider the problem of computing the Wasserstein barycenter with support size at most n for the $p = 2$ objective. There exists a matrix $A \in \mathbb{R}^{nk \times d}$ and a set S of rank n orthogonal projection matrices in $\mathbb{R}^{nk \times nk}$ such that the first problem is equivalent to computing*

$$\boldsymbol{P}^* = \operatorname*{argmin}_{\boldsymbol{P} \in S} \|\boldsymbol{A} - \boldsymbol{P}\boldsymbol{A}\|_F^2.$$

The proof of Theorem 6.1 is given in Section E.

We also prove the following NP hardness result in Section F which complements the hardness results in [AB21, BP21].

Theorem 6.2. *It is NP-hard to approximate an optimal Wasserstein barycenter of fixed support size up to a multiplicative factor* 1.0013.

7 Experiments

In this section, we empirically verify that dimensionality reduction can provide large computational savings without significantly reducing accuracy. We use the following datasets in our experiments.

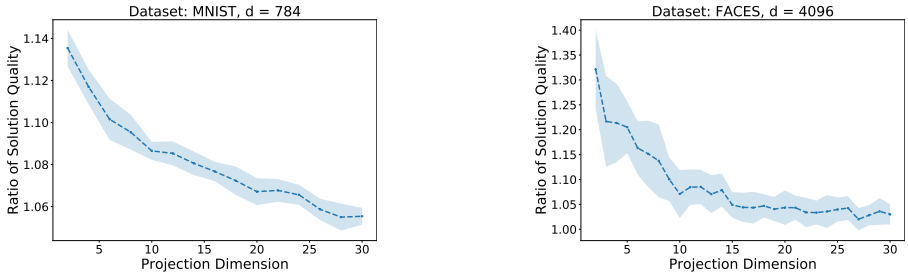
FACES dataset: This dataset is used in the influential ISOMAP paper and consists of 698 images of faces in dimension 4096 [TSL00]. We form $k = 2$ distributions by splitting the images facing to the "left" versus the ones facing "right." This results in ~ 350 uniform point masses per distribution.

MNIST dataset: We subsample 10^4 images from the MNIST test dataset (dimension 784). We split the images by their digit class which results in $k = 10$ distributions with $\sim 10^3$ uniform point masses each in $\mathbb{R}^{784}$.

Experimental setup. We project our datasets in dimensions d ranging from $d = 2$ to $d = 30$ and compute the Wasserstein barycenter for $p = 2$. For FACES, we limit the support size of the barycenter to be at most 5 points in $\mathbb{R}^{4096}$ (since the barycenter should intuitively return an "interpolation" between the left and right facing faces, it should not be supported on too many points). For MNIST we limit the support size of the barycenter to be at most 40. We then take the barycenter found in the lower dimension and compare its cost in the higher dimension (see Algorithm 1) against the Wasserstein barycenter found in the higher dimension.

We use the code and default settings from [Ye19] to compute the Wasserstein barycenter; this implementation has been applied in previous empirical papers [YWWL17]. While we fix this implementation, note that dimensionality reduction is extremely flexible and can work with any algorithm or implementation (see Algorithm 1) and we would expect it to produce similar results.

Results. Our results are displayed in Figure 2. We see that for both datasets, reducing the dimension to $d = 30$ only increases the cost of the solution by 5%. This is **1-2** orders of magnitude smaller than from the original dimensions of 784 and 4096 for MNIST and FACES respectively. The average time taken to run the Wasserstein barycenter computation algorithm in $d = 30$ was 73% and 9% of the time taken to run in the full dimensions respectively.



(a) MNIST (b) Faces

Figure 2: Ratio of the quality of solution found in the lower dimension versus the original dimension. Result displays average of 20 independent trials and ± 1 standard deviation is shaded.

Coreset experiments. Our coreset result reduces the number of distributions k through sensitivity (importance) sampling. We created a synthetic dataset with large k but small n and d to emphasize the advantage of sensitivity sampling over uniform sampling. We have $k = 50,000$ distributions that each consists of a single point mass in $\mathbb{R}$. The first $k - 1$ distributions are all supported at the origin while one distribution is supported at $x = k$. We consider the $p = 2$ case and limit the support size of the barycenter to also be 1. Let $\text{cost}_{\text{orig}}(\nu)$ denote the cost of ν on the original objective (1) and let $\text{cost}_{\text{core}}(\nu)$ the cost of (1) when evaluated on a coreset. We record the relative error $|\text{cost}_{\text{core}}(\nu) - \text{cost}_{\text{orig}}(\nu)|/|\text{cost}_{\text{orig}}(\nu)|$ evaluated at $\nu = \delta_x$, i.e. a single unit point mass at x, for $x = 0, 1, 10$. We then average the results across 10 trials each. As x (the point on which the query distribution is supported) grows bigger, the associated cost became bigger, hence decreasing the relative error. Other query locations displayed the same trend. See Figure 3 for more details.

Method	# of samples	% error at query			
		$x = 100$	$x = 10$	$x = 1$	$x = 0$
Uniform sampling	1000	0.986	9.087	49.998	100
Sensitivity sampling	10	0.0040	0.0036	0.0020	0

Figure 3: Even with much fewer samples, sensitivity sampling outperforms uniform sampling for a number of query locations, averaged across 10 repetitions.

Acknowledgments

Sandeep Silwal was supported in part by a NSF Graduate Research Fellowship Program. Samson Zhou was supported by a Simons Investigator Award of David P. Woodruff.

References

[AB99] Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

[AB21] Jason M. Altschuler and Enric Boix-Adserà. Wasserstein barycenters are np-hard to compute. *CoRR*, abs/2101.01100, 2021.

[ABA21] Jason M Altschuler and Enric Boix-Adsera. Wasserstein barycenters can be computed in polynomial time in fixed dimension. *Journal of Machine Learning Research*, 22(44):1–19, 2021.

[ABM16] Ethan Anderes, Steffen Borgwardt, and Jacob Miller. Discrete wasserstein barycenters: Optimal transport for discrete data. *Mathematical Methods of Operations Research*, 84, 10 2016.

[AC09] Nir Ailon and Bernard Chazelle. The fast johnson–lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, 39(1):302–322, 2009.

[AC11a] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM J. Math. Analysis*, 43:904–924, 01 2011.

[AC11b] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[ÁDCM16] Pedro C Álvarez-Esteban, E Del Barrio, JA Cuesta-Albertos, and C Matrán. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

[BBC+19] Luca Becchetti, Marc Bury, Vincent Cohen-Addad, Fabrizio Grandoni, and Chris Schwiegelshohn. Oblivious dimension reduction for k-means: beyond subspaces and the johnson-lindenstrauss lemma. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC*, pages 1039–1050, 2019.

[BCC+15] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[BLK17] Olivier Bachem, Mario Lucic, and Andreas Krause. Practical coreset constructions for machine learning, 2017.

[BP21] Steffen Borgwardt and Stephan Patterson. On the computational complexity of finding a sparse wasserstein barycenter. *J. Comb. Optim.*, 41(3):736–761, 2021.

[BT97] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

[CC06] Miroslav Chlebík and Janka Chlebíková. Complexity of approximating bounded variants of optimization problems. *Theor. Comput. Sci.*, 354(3):320–338, 2006.

[CD14] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693, 2014.

[CEM+15] Michael B. Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC*, pages 163–172, 2015.

[CG15] John P Cunningham and Zoubin Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *The Journal of Machine Learning Research*, 16(1):2859–2900, 2015.

[CMRS20] Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J Stromme. Gradient descent algorithms for bures-wasserstein barycenters. In *Conference on Learning Theory*, pages 1276–1304, 2020.

[COO15] Guillaume Carlier, Adam Oberman, and Edouard Oudet. Numerical methods for matching for teams and wasserstein barycenters. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1621–1642, 2015.

[EHJK20] Filip Elvander, Isabel Haasler, Andreas Jakobsson, and Johan Karlsson. Multi-marginal optimal transport using partial information with applications in robust localization and sensor fusion. *Signal Processing*, 171:107474, 2020.

[Fel20] Dan Feldman. Introduction to core-sets: an updated survey. *CoRR*, abs/2011.09384, 2020.

[HNY+17] Nhat Ho, XuanLong Nguyen, Mikhail Yurochkin, Hung Hai Bui, Viet Huynh, and Dinh Phung. Multilevel clustering via wasserstein means. In *International Conference on Machine Learning*, pages 1501–1509. PMLR, 2017.

[IN07] Piotr Indyk and Assaf Naor. Nearest-neighbor-preserving embeddings. *ACM Trans. Algorithms*, 3(3):31–es, August 2007.

[JL84] William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

[Kir34] M. Kirszbraun. Über die zusammenziehende und lipschitzsche transformationen. *Fundamenta Mathematicae*, 22(1):77–108, 1934.

[KMN11] Daniel M. Kane, Raghu Meka, and Jelani Nelson. Almost optimal explicit johnson-lindenstrauss families. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques - 14th International Workshop, APPROX, and 15th International Workshop, RANDOM. Proceedings*, pages 628–639, 2011.

[KTD+19] Alexey Kroshnin, Nazarii Tupitsa, Darina Dvinskikh, Pavel Dvurechensky, Alexander Gasnikov, and Cesar Uribe. On the complexity of approximating wasserstein barycenters. In *International conference on machine learning*, pages 3530–3540, 2019.

[Lan18] Harry Lang. *Streaming Coresets for High Dimensional Geometry*. PhD thesis, Johns Hopkins University, 7 2018.

[LFKF18] Mario Lucic, Matthew Faulkner, Andreas Krause, and Dan Feldman. Training gaussian mixture models at scale via coresets. *Journal of Machine Learning Research*, 18(160):1–25, 2018.

[LHC+20] Tianyi Lin, Nhat Ho, Xi Chen, Marco Cuturi, and Michael I Jordan. Fixed-support wasserstein barycenters: Computational hardness and fast algorithm. *Advances in Neural Information Processing Systems*, 33, 2020.

[LN16] Kasper Green Larsen and Jelani Nelson. The johnson-lindenstrauss lemma is optimal for linear dimensionality reduction. In *43rd International Colloquium on Automata, Languages, and Programming, ICALP*, 2016.

[LN17] Kasper Green Larsen and Jelani Nelson. Optimality of the johnson-lindenstrauss lemma. In *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS*, pages 633–638, 2017.

[LSPC19] Giulia Luise, Saverio Salzo, Massimiliano Pontil, and Carlo Ciliberto. Sinkhorn barycenters with free support via frank-wolfe algorithm. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems*, pages 9318–9329, 2019.

[LSW17] Euiwoong Lee, Melanie Schmidt, and John Wright. Improved and simplified inapproximability for k-means. *Inf. Process. Lett.*, 120:40–43, 2017.

[MC19] Boris Muzellec and Marco Cuturi. Subspace detours: Building transport plans that are optimal on subspace projections. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS*, pages 6914–6925, 2019.

[MMR19] Konstantin Makarychev, Yury Makarychev, and Ilya P. Razenshteyn. Performance of johnson-lindenstrauss transform for k-means and k-medians clustering. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing, STOC*, pages 1027–1038, 2019.

[PW09] Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th international conference on computer vision*, pages 460–467. IEEE, 2009.

[RGT97] Yossi Rubner, Leonidas J Guibas, and Carlo Tomasi. The earth mover's distance, multi-dimensional scaling, and color-based image retrieval. In *Proceedings of the ARPA image understanding workshop*, volume 661, page 668, 1997.

[RPDB11] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446, 2011.

[RU02] Ludger Rüschendorf and Ludger Uckelmann. On the n-coupling problem. *Journal of multivariate analysis*, 81(2):242–258, 2002.

[SCSJ17] Matthew Staib, Sebastian Claici, Justin Solomon, and Stefanie Jegelka. Parallel streaming wasserstein barycenters. pages 2647–2658, 2017.

[SDGP+15] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):1–11, 2015.

[SLD18] Sanvesh Srivastava, Cheng Li, and David B Dunson. Scalable bayes via barycenter in wasserstein space. *The Journal of Machine Learning Research*, 19(1):312–346, 2018.

[TSL00] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[Vil08] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[Wai19] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[Ye19] Jianbo Ye. Wbc-matlab. https://github.com/bobye/WBC_Matlab, 2019.

[YWWL17] Jianbo Ye, P. Wu, J. Z. Wang, and Jia Li. Fast discrete distribution clustering using wasserstein barycenter with sparse support. *IEEE Transactions on Signal Processing*, 65:2317–2332, 2017.

A Proofs for Section 3

Theorem A.1. *Let $p \neq 2$. Consider a JL projection π from $\mathbb{R}^d$ to $\mathbb{R}^m$ for $m = O(\log(nk/\delta)p^2/\varepsilon^2)$. Then we have*

$$\mathbb{P}\left(\text{cost}_p(\pi S) \in [(1-\varepsilon)\cdot \text{cost}_p(S), (1+\varepsilon)\cdot \text{cost}_p(S)] \textit{ for all solutions } S\right) \geq 1-\delta,$$

where the probability is taken over the randomness in the projection π.

Proof of Theorem A.1. We again assume that the distances between point masses among all the nk points in the distributions $\mu_1, \ldots, \mu_k$ are preserved up to $1 \pm \varepsilon$. By Theorem 3.2, dimensionality reduction gives us a $(1+\varepsilon)$-Lipschitz map $\widetilde{\pi} : \mathbb{R}^d \to \mathbb{R}^m$ as well as $\widetilde{\varphi} : \mathbb{R}^m \to \mathbb{R}^d$.

Now consider an arbitrary solution $S = (S_1, \ldots, S_n)$. We first show that $\text{cost}(S) \leq (1+\varepsilon)^p \text{cost}(\pi S)$ where $\text{cost}(\pi S)$ is the cost of the solution S evaluated in the projected space. Indeed for any S_j, the objective in the original dimension $\mathbb{R}^d$ is

$$\sum_{x \in S_j} w_j(x)\|x - \nu^j\|^p.$$

Let ν^j denote the argmin of this objective in $\mathbb{R}^d$ and let u^j denote the argmin for this same objective but in the projected space $\mathbb{R}^m$, i.e.

$$u^j = \operatorname*{argmin}_{u \in \mathbb{R}^m} \sum_{x \in S_j} w_j(x)\|\pi x - u\|^p.$$

Then we have

$$\begin{aligned}
\sum_{x \in S_j} w_j(x)\|x - \nu^j\|^p &\leq \sum_{x \in S_j} w_j(x)\|x - \widetilde{\varphi}(u^j)\|^p && (\nu^j \text{ is more optimal than } \widetilde{\varphi}(u)) \\
&= \sum_{x \in S_j} w_j(x)\|\widetilde{\varphi}(\pi x) - \widetilde{\varphi}(u^j)\|^p && (\widetilde{\varphi} = \pi^{-1} \text{ for } x \in S_j) \\
&\leq \sum_{x \in S_j} w_j(x)(1+\varepsilon)^p\|\pi x - u^j\|^p && (\widetilde{\varphi} \text{ is } (1+\varepsilon)\text{-Lipschitz}) \\
&= (1+\varepsilon)^p \sum_{x \in S_j} w_j(x)\|\pi x - u^j\|^p.
\end{aligned}$$

Summing over all S_j finally leads to $\text{cost}(S) \leq (1+\varepsilon)^p \text{cost}(\pi S)$. A similar reasoning also gives $\text{cost}(\pi S) \leq (1+\varepsilon)^p \text{cost}(S)$ and combining these two statements and adjusting ε proves the theorem. □

B Proofs for Section 4

In this section, we give the missing proofs from Section 4. Our main goal will be to prove Theorem 4.1; we also describe a "faster" dimension reduction map at the end of the section. To prove Theorem 4.1, we will actually first prove a version of the theorem with a slightly rescaled value of ε (Theorem B.13). Theorem 4.1 follows immediately by "undoing" the rescaling.

We adapt this analysis to the Wasserstein barycenter problem by handling four additional issues: (i) the input points are weighted since they come from probability distributions; (ii) input points may be assigned to multiple support points in the barycenter; (iii) each barycenter point is constrained to receive a specific amount of mass under optimal transport; and (iv) the distorted points must not contribute large error to the cost induced by the Wasserstein barycenter. Issues (i) and (ii) are problematic because previous structural results for the distortion graph do not rule out a large weighted fraction of the distances being distorted. Furthermore, issues (iii) and (iv) are problematic because we cannot isolate each point in a probability distribution to a specific barycenter. We again consider a hypothetical distortion graph on the $k \cdot \text{poly}(n)$ points in $\mathbb{R}^d$ with nonzero support in the k distributions and connect an edge between each pair of points if their pairwise distance is distorted by the random projection map π by at least a $(1+\varepsilon)$-factor. To resolve issue (1), we give a combinatorial

argument that shows that the distortion graph for π is everywhere sparse for a weighted notion of sparsity. To resolve issues (2) and (3), we define a mapping for each point in a probability distribution that partitions its mass among the barycenters. Using the everywhere-sparse distortion graph, we show a robust 1-point extension theorem that the pairwise distances from the barycenter to a large *weighted* fraction of the points is preserved. Finally to resolve issue (4), we show that the remaining weighted fraction of points incurs a cost that is at most ε-fraction of the optimal cost induced by the Wasserstein barycenter.

The structure of the proof is as follows. To prove that the cost of any solution (in the sense of Definition 2.1) is preserved, we first show that the cost of the flow from a weighted cluster of points in the μ_is to one particular support point in the barycenter is preserved (Theorem B.12). This in turn rests on the fact that weighted cluster costs are preserved when only a small weighted fraction of the cluster distances are distorted (Theorem B.1).

In summary, the overall proof structure is

$$\text{Theorem B.1} \implies \text{Theorem B.12} \implies \text{Theorem B.13} \implies \text{Theorem 4.1}.$$

We begin by proving Theorem B.1, which should be considered the weighted analog to Theorem 3.3 in [MMR19].

Theorem B.1. *Let $X \subset \mathbb{R}^d$ be a finite set of weighted points and the map $\phi : X \to \mathbb{R}^m$ have a distortion graph G for X that is α-sparse (with respect to the weight of X), with $\alpha \leq 1/10^{p+1}$. Then for every $p \geq 1$ and $D = (1+\varepsilon)^p(1 + 3^{p+2} + \alpha^{1/(p+1)})$, we have*

$$\frac{1}{D}\text{cost}_p(X) \leq \text{cost}_p(\phi(X)) \leq D\text{cost}_p(X)$$

where cost_p is the cost of solving a clustering on X with only 1 center under the cost function $\|\cdot\|_2^p$.

Unfortunately, the results of [MMR19] do not immediately imply the corresponding sparsity results for weighted graphs. For example, a vertex that has edges to a small fraction of its neighbors may still have an edge to a large weighted fraction of its neighbors. Thus we show the weighted analogs of the structural results from [MMR19]. The following lemma is analogous to Lemma 4.1 in [MMR19], extending the properties to handle weighted sets X.

Lemma B.2. *Let X be a finite set and $V \subset X$ be a random subset of X. Let $\alpha \in (0, 1/2)$ and suppose that $\mathbf{Pr}[x \in V] \geq 2\alpha$ for each $x \in X$. Then there exist a random set $R \subset V$ and a deterministic measure μ on X such that*

1. *$\mu(x) \geq \frac{w(x)}{w(V \setminus R)}$ for every $x \in V \setminus R$*
2. *$\mathbf{Pr}[x \in R] \leq 2\alpha$ for every $x \in X$*
3. *$\mu(X) = \sum_{x \in X} \mu(x) \leq \frac{\mathbf{Pr}[V \neq \emptyset]}{\alpha^2}$*

Proof. Since X is a finite set, we truncate (or discretize) the weights of the elements in X and without loss of generality suppose that there exists a sufficiently large integer $N > 0$ such that for each $x \in X$, there exists some integer $i \leq N$ such that $w(x) = \frac{i}{N}$. We then prove our claim by induction on the weight of the set X. If $w(X) = 0$ so that X is empty, then the claim trivially holds. Now we suppose that $w(X) = \frac{k}{N}$ and the statement holds for all sets X' with weight $w(X') = \frac{k'}{N}$, where $k' < k$ are non-negative integers; we show the statement holds for X.

Let $\ell = \alpha \cdot w(X)$ and define a deterministic set X' and a random subset $V' \subset X'$ by:

$$X' = \{x : \mathbf{Pr}[x \in V \text{ and } w(V) < \ell] \geq 2\alpha\}$$
$$V' = \begin{cases} V \cap X', & \text{if } w(V) < \ell \\ \emptyset, & \text{otherwise} \end{cases}$$

We first show that there exists an $x_0 \in X$ such that $\mathbf{Pr}[x_0 \in V \text{ and } w(V) < \ell] \leq \alpha$, which implies that $x_0 \notin X'$ and thus $w(X') < w(X)$. We show that the average value of $\mathbf{Pr}[x \in V \text{ and } w(V) < \ell]$

for $x \in X$ is at most α, which implies the existence of such an x_0. Since $w(V) \cdot \mathbb{1}\{w(V) < \ell\}$ is always at most ℓ, then

$$\frac{1}{w(X)} \sum_{x \in X} \mathbf{Pr}\left[x \in V \text{ and } w(V) < \ell\right] = \frac{1}{w(X)} \sum_{x \in X} \mathbb{E}\left[\mathbb{1}\{x \in V \text{ and } w(V) < \ell\}\right]$$
$$= \frac{1}{w(X)} \mathbb{E}\left[\sum_{x \in X} \mathbb{1}\{x \in V \text{ and } w(V) < \ell\}\right] = \frac{1}{w(X)} \mathbb{E}\left[w(V) \cdot \mathbb{1}\{w(V) < \ell\}\right] \leq \frac{\ell}{w(X)} = \alpha.$$

Because $w(X') < w(X)$ and $\mathbf{Pr}\left[x \in V'\right] = \mathbf{Pr}\left[x \in V \text{ and } w(V) < \ell\right] \geq 2\alpha$ for each $x \in X'$ by definition of X', then we apply the inductive hypothesis to X' and V'. Hence, there exist a random set $R' \subset X'$ and a measure μ' on X' such that the above claims 1-3 hold for X' and V'. We then define a measure μ on X and random subset $R \subset V$ by:

$$\mu(x) = \begin{cases} \mu'(x) + \frac{w(x)}{\ell}, & \text{if } x \in X' \\ \frac{w(x)}{\ell}, & \text{otherwise.} \end{cases}$$
$$R = \begin{cases} R' \cup (V \setminus X'), & \text{if } w(V) < \ell \\ R', & \text{otherwise.} \end{cases}$$

We claim that R and μ satisfy the desired properties.

Property 1: $\mu(x) \geq \frac{w(x)}{w(V \setminus R)}$ for each $x \in V \setminus R$. Let $x \in V \setminus R$.
We have three possible cases. (1) If $x \in X'$ and $w(V) < \ell$, then $V \setminus R = V' \setminus R'$ by the definition of R. Hence, $\mu(x) > \mu'(x) \geq \frac{w(x)}{w(V' \setminus R)} = \frac{w(x)}{w(V \setminus R)}$ by the inductive hypothesis. (2) If $x \in X'$ and $w(V) \geq \ell$, then $\mu(x) \geq \frac{w(x)}{\ell} \geq \frac{w(x)}{w(V)}$. Since $w(V) \geq \ell$, we also have $V' = \emptyset$, so by the definition of R, we have $R = R' \subset V' = \emptyset$. Thus, $\frac{w(x)}{w(V)} = \frac{w(x)}{w(V \setminus R)}$ so that $\mu(x) \geq \frac{w(x)}{w(V \setminus R)}$. (3) If $x \notin X'$, then $x \in V \setminus X' \subset R' \cup (V \setminus X')$. Since $X \in V \setminus R$, then $x \notin R$. Thus, $R \neq R' \cup (V \setminus X')$. By the definitions of μ and R, we have that $w(V) \geq \ell$ and $\mu(x) = \frac{w(x)}{\ell}$, so that $\mu(x) \geq \frac{w(x)}{w(V)}$. Since $w(V) \geq \ell$, then $\mu(x) \geq \frac{w(x)}{w(V)} = \frac{w(x)}{w(V \setminus R)}$.

Property 2: $\mathbf{Pr}\left[x \in R\right] \leq 2\alpha$. If $x \in X'$, then by the inductive hypothesis, $\mathbf{Pr}\left[x \in R\right] = \mathbf{Pr}\left[x \in R'\right] \leq 2\alpha$. If $x \notin X'$, then by the definitions of R and X' respectively, we have that $\mathbf{Pr}\left[x \in R\right] = \mathbf{Pr}\left[x \in V \text{ and } w(V) < \ell\right] \leq 2\alpha$.

Property 3: $\mu(X) = \sum_{x \in X} \mu(x) \leq \mathbf{Pr}\left[V \neq \emptyset\right]/\alpha^2$. By the inductive hypothesis, $\mu'(X') \leq \mathbf{Pr}\left[V' \neq \emptyset\right]/\alpha^2$. Therefore, $\mu(X) = \mu'(X') + \frac{w(X)}{\ell} \leq \frac{\mathbf{Pr}\left[V' \neq \emptyset\right]}{\alpha^2} + \frac{1}{\alpha}$. Note that $\mathbf{Pr}\left[V \neq \emptyset\right] - \mathbf{Pr}\left[V' \neq \emptyset\right] \geq \alpha$ implies $\frac{\mathbf{Pr}\left[V' \neq \emptyset\right]}{\alpha^2} + \frac{1}{\alpha} \leq \frac{\mathbf{Pr}\left[V \neq \emptyset\right]}{\alpha^2}$, which would imply the desired claim; thus it suffices to prove $\mathbf{Pr}\left[V \neq \emptyset\right] - \mathbf{Pr}\left[V' \neq \emptyset\right] \geq \alpha$.

Observe that if $w(V) \geq \ell$, then $V \neq \emptyset$ but $V' = \emptyset$. Since $V' \subset V$, then

$$\mathbf{Pr}\left[V \neq \emptyset\right] - \mathbf{Pr}\left[V' \neq \emptyset\right] = \mathbf{Pr}\left[V \neq \emptyset \text{ and } V' = \emptyset\right] \geq \mathbf{Pr}\left[w(V) \geq \ell\right].$$

Recall that we previously showed the existence of an $x_0 \in X$ with $\mathbf{Pr}\left[x_0 \in V \text{ and } w(V) < \ell\right] \leq \alpha$. On the other hand, $\mathbf{Pr}\left[x \in V\right] \geq 2\alpha$ for all $x \in X$. Thus,

$$\mathbf{Pr}\left[w(V) \geq \ell\right] \geq \mathbf{Pr}\left[x_0 \in V \text{ and } w(V) \geq \ell\right] \geq \mathbf{Pr}\left[x_0 \in V\right] - \mathbf{Pr}\left[x_0 \in V \text{ and } w(V) < \ell\right] \geq \alpha.$$

Hence, we have shown that all three properties are satisfied by R and μ, which completes the induction for X. □

The following claim is analogous to Corollary 4.2 in [MMR19], again extending the properties to handle weighted sets X.

Corollary B.3. *Let X be a finite set, $V \subset X$ be a random subset of X, and $\alpha \in (0, 1/2)$. Then there exist a random set $R \subset V$ and a measure μ on X such that*

1. $\mu(x) \geq \frac{w(x)}{w(V \setminus R)}$ *for every* $x \in V \setminus R$
2. $\mathbf{Pr}[x \in R] \leq 2\alpha$ *for every* $x \in X$
3. $\mu(X) = \sum_{x \in X} \mu(x) \leq \frac{1}{\alpha^2}$

Proof. Let $X' = \{x : \mathbf{Pr}[x \in V] \geq 2\alpha$. By applying Lemma B.2 to X' and $V' = V \cap X'$ and set

$$R = R' \cup (V \setminus X'),$$
$$\mu(x) = \begin{cases} \mu'(x), & \text{if } x \in X' \\ 0, & \text{otherwise} \end{cases}.$$

□

We also have the following analog to Observation 4.3 in [MMR19].

Observation B.4. *Let R be defined as in Corollary B.3 and $V_0 = V \setminus R$. Then for every $S \subset V_0$, we have $w(S) \leq \mu(S) \cdot w(V_0)$.*

Proof. By Corollary B.3, we have $\mu(x) \geq w(x)/w(V_0)$ for every $x \in S$. Thus, $\mu(S) \geq w(S)/w(V_0)$, so $w(S) \leq \mu(S) \cdot w(V_0)$. □

The following is analogous to Theorem 3.2 in [MMR19].

Theorem B.5 (Theorem 3.2 in [MMR19])**.** *Consider a finite set X and a random graph $H = (V, E)$, where V is a random subset of X and E is a random set of edges between vertices in V (there are no independence assumptions or any other implicit assumptions about the distribution of V and E). Let $\alpha \in (0, 1/2)$. Assume that $\mathbf{Pr}[(x, y) \in E] \leq \delta \leq \alpha$ for every $x, y \in X$. Then there exists a random subset $V' \subset V$ such that*

- *$H[V']$ is α-everywhere sparse,*
- *$\mathbf{Pr}[u \in V \setminus V'] \leq 600\delta/\alpha^6$ for all $u \in X$.*

Proof. Let $\beta = \alpha/(1 + \alpha)$ and $\alpha' = \alpha/3$. Applying Corollary B.3 with α', we get a deterministic measure μ on V and a random set $R \subset V$. Consider the canonical product measure

$$\mu^{\otimes 2}((x, y)) = \mu(x)\mu(y)$$

for all $x, y \in X$. We define V' according to the measure $\mu^{\otimes 2}$.

Case 1: $\mu^{\otimes 2}(E) \geq \beta^2$. In this case, we let $V' = \emptyset$.

Case 2: $\mu^{\otimes 2}(E) < \beta^2$. In this case, define $V_0 = V \setminus R$. We say that $x \in X$ is bad if $\mu(\{y \in V_0 : (x, y) \in E\}) \geq \beta$. Let B denote the set of bad vertices and define

$$V' = V \setminus (R \cup B) = V_0 \setminus B.$$

Our goal is to verify that in both of the above cases, both of the desired properties of Theorem B.5 hold. First we handle the easier Case 1. There, the graph $H[V']$ is empty so the conclusion trivially holds. For the second condition, note that

$$\mathbf{Pr}\left[\mu^{\otimes 2}(E) \geq \beta^2\right] \leq \frac{\mathbb{E}\left[\mu^{\otimes 2}(E)\right]}{\beta^2} \leq \frac{\delta M^2}{\beta^2}$$

where $M = \mu(X)$ and we have used the fact that $\mathbf{Pr}[(x, y) \in E] \leq \delta$. Then from our choice of β and α' (fill in details in a bit), the above probability is at most $C\alpha$.

We now verify the second case. First, we check that $H[V']$ is α-everywhere sparse. This is equivalent to checking that the weighted degree of every vertex x in $H[V']$ is at most α fraction of the total weight. That is, we need to check:

$$w(\{y \in V' : (x, y) \in E\}) \leq \alpha w(V').$$

Analogous to the proof of Theorem 3.2 in [MMR19], we have $\mu^{\otimes 2}(E) \geq \beta\mu(B)$ and therefore, $\mu(V_0 \cap B) \leq \mu(B) \leq \beta$ since we are in Case 2. Now for $x \in V'$, we similarly have $\mu(y \in V' : (x, y) \in V') \leq \beta$. Combining the above findings with Observation B.4, we conclude the following two statements:

- $w(\{y \in V' : (x, y) \in E\}) \leq \beta w(V_0)$,
- $w(V_0 \cap B) \leq \beta w(V_0)$.

From the second relation, we have $w(V') = w(V_0 \setminus B) \geq (1 - \beta)w(V_0)$ and using the first relation, we can conclude that

$$w(\{y \in V' : (x, y) \in E\}) \leq \frac{\beta}{1-\beta} w(V').$$

Finally, the same probability bound as in the end of the proof of Theorem 3.2 in [MMR19] allows us to say that in Case 2, the probability of $x \in B$ is at most $\delta M/\beta$. Finally, combining all the probabilities from Case 1 and Case 2, we can conclude identically as in Theorem 3.2 in [MMR19] that $\mathbf{Pr}\left[u \in V\ V'\right] \leq \delta M^2/\beta^2 + 2\alpha' + \delta M/\beta \leq 600\delta/\alpha^6$. □

We now state an analogous version of Theorem 5.2 from [MMR19] that is suitable for our purposes.

Theorem B.6 (Robust Kirszbraun Theorem)**.** *Consider two finite (multi) sets of points $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}^m$ and a map $\varphi : X \to Y$. Let $G = (X, E)$ be the distance expansion graph for φ with respect to the Euclidean distance with vertex weights given by $w : X \to \mathbb{R}^{\geq 0}$. Suppose that G is α-everywhere sparse according to Definition B.10. Then for every $u \in \mathbb{R}^d$ and $\varepsilon > 0$, there exists $v \in \mathbb{R}^m$ and $X' \subset X$ such that for all $x \in X \setminus X'$,*

$$\|\phi(x) - v\| \leq (1 + \varepsilon)\|x - u\|,$$

and $w(X') \leq \alpha'(\varepsilon)w(X)$ where $\alpha'(\varepsilon) = 2(1 + \varepsilon)^2\alpha/\varepsilon$.

The proof of Theorem 5.2 from [MMR19] carries over in a straightforward fashion to the proof of Theorem B.6 above. In particular, we just outline the small changes that need to occur to carry the proof over.

Proof Sketch. In [MMR19], the following polytope is defined:

$$\Lambda_\eta = \{\lambda \in \mathbb{R}^X : \sum_{x \in X} \lambda_x = 1; 0 \leq \lambda_{x'} \leq \eta \text{ for all } x' \in X\}$$

for $\eta = (\alpha'(\varepsilon)n)^{-1}$. For us, we define a slightly modified polytope which includes the weights of elements of X:

$$\Lambda_\eta = \{\lambda \in \mathbb{R}^X : \sum_{x \in X} \lambda_x = 1; 0 \leq \lambda_{x'} \leq w(x')\eta \text{ for all } x' \in X\}$$

where $\eta = (\alpha'(\varepsilon)w(X))^{-1}$. Then for every $\lambda \in \Lambda_\eta$, $u' \in \mathbb{R}^{d'}$, and $v' \in R^{d''}$, we similarly let

$$f(X, \lambda, u') = \sum_{x \in X} \lambda_x \|u' - x\|^2 \text{ and } f(\phi(X), \lambda, v') = \sum_{x \in X} \lambda_x \|v' - \phi(x)\|^2.$$

Since Λ_η is still a convex polytope, we recover the statement

$$\max_{v' \in \mathbb{R}^{d''}} \min_{\lambda \in \Lambda_\eta} F(v', \lambda) \geq 0 \tag{5}$$

where $F(v', \lambda) = (1 + \varepsilon)^2 f(X, \lambda, u) - f(\varphi(X), \lambda, v')$. Now to finish the rest of the proof, let v be the point that maximizes the functional $\min_{\lambda \in \Lambda_\eta} F(v, \lambda)$. By (5), we know that $F(v, \lambda) \geq 0$ for all $\lambda \in \Lambda_\eta$. Now consider the set

$$S = \{x \in X : \|\varphi(x) - v\| \geq (1 + \varepsilon)\|x - u\|\}.$$

If $S = \emptyset$, we are done so otherwise, define λ^* as

$$\lambda^*_x = \begin{cases} \frac{w(x)}{w(S)} & \text{if } x \in S, \\ 0 & \text{otherwise .} \end{cases}$$

By the definition of S, we have $(1+\varepsilon)\|x-u\|^2 - \|\varphi(x)-v\|^2 < 0$ which implies

$$F(v,\lambda^*) = \frac{1}{w(S)}\sum_{x\in S} w(x)\left((1+\varepsilon)\|x-u\|^2 - \|\varphi(x)-v\|^2\right) < 0$$

and thus, $\lambda^* \notin \Lambda_\eta$. Therefore, $1/w(S) > \eta$ and $w(S) < 1/\eta = \alpha'(\varepsilon)w(X)$. This finishes the proof of Theorem B.6. □

We also need the following analogous version of Lemma 5.3 in [MMR19]. The proof differs in that we have to consider a careful weighting scheme whereas in [MMR19] it was more straightforward. We first require the following property:

Lemma B.7 (Lemma A.1 in [MMR19])**.** *Let x and $y_1,\ldots,y_r$ be non-negative real numbers, and $\varepsilon > 0$, $p \geq 1$. Then*

$$\left(x + \sum_{i=1}^{r} y_i\right)^p \leq (1+\varepsilon)^{p-1}x^p + \left(\frac{(1+\varepsilon)r}{\varepsilon}\right)^{p-1}\sum_{i=1}^{r} y_i^p.$$

Lemma B.8 (Lemma 5.3 in [MMR19])**.** *Consider two finite multisets of points $X \subset \mathbb{R}^d$ and $Y \subset \mathbb{R}^m$ of the same size and a one-to-one map $\varphi : X \to Y$. Let $G = (X, E)$ be the distance expansion graph for φ with respect to the Euclidean distance with a weight function $w : X \to \mathbb{R}^{\geq 0}$. Suppose that G is α-everywhere sparse with $\alpha \leq 1/10^{p+1}$. Then, for every $p \geq 1$, we have the following inequality on the cost of the clusters X and Y (with the same weights as w)*

$$\mathrm{cost}_p(Y) \leq (1 + 3^{p+2}\alpha^{1/(p+1)})\mathrm{cost}_p(X).$$

Proof. Let $\varepsilon = \alpha^{1/(p+1)}$ and let u^* be the optimal center for the cluster X. By Theorem B.6, there exists a set $\widetilde{X} \subset X$ and a point $v^* \in \mathbb{R}^m$ such that for $x \in \widetilde{X}$,

$$\|\varphi(x) - v^*\| \leq (1+\varepsilon)\|x-u^*\|$$

where $w(X \setminus \widetilde{X}) \leq \alpha' w(X)$ with $\alpha' \leq 2(1+\varepsilon)^2\alpha/\varepsilon$. By definition, it follows that

$$\begin{aligned}
\mathrm{cost}_p(Y) &\leq \sum_{y\in Y} w(y)\|y-v^*\|^p \\
&= \sum_{x\in X} w(x)\|\varphi(x)-v^*\|^p \ (x \text{ and } \varphi(x) = y \text{ have the same weight}) \\
&= \sum_{x\in \widetilde{X}} w(x)\|\varphi(x)-v^*\|^p + \sum_{x\notin \widetilde{X}} w(x)\|\varphi(x)-v^*\|^p \\
&\leq (1+\varepsilon)^p \sum_{x\in \widetilde{X}} w(x)\|x-u^*\|^p + \sum_{x\notin \widetilde{X}} w(x)\|\varphi(x)-v^*\|^p.
\end{aligned}$$

We now try to bound $\|\varphi(x)-v^*\|^p$ for $x \notin \widetilde{X}$. We will bound this quantity using a slightly stronger claim that applies for all $x \in X$. Indeed, fix an arbitrary x and consider the set I_x of its non neighbors in the distance expansion graph G. Note that $w(I_x) \geq (1-\alpha)w(X)$ and thus, $w(I_x \cap \widetilde{X}) \geq (1-\alpha-\alpha')w(X) > 0$ if the total weight $w(X)$ is positive. Consider an arbitrary $x' \in I_x \cap \widetilde{X}$. Then it follows that

$$\begin{aligned}
\|\varphi(x)-v^*\| &\leq \|\varphi(x)-\varphi(x')\| + \|\varphi(x')-v^*\| \\
&\leq (1+\varepsilon)\|x-x'\| + (1+\varepsilon)\|x'-u^*\| \ (\text{using the fact that } x' \in I_x \cap \widetilde{X}) \\
&= (1+\varepsilon)\|x-u^*\| + (2+2\varepsilon)\|x'-u^*\|.
\end{aligned}$$

Applying Lemma B.7, we have that

$$\|\varphi(x)-v^*\| \leq (1+\varepsilon)^p\|x-u^*\|^p + \frac{3^p}{\varepsilon^{p-1}}\|x'-u^*\|^p \tag{6}$$

for ε sufficiently small. We now average equation (6) over all possible x'. This gives us

$$\begin{aligned}\|\varphi(x) - v^*\| &\leq (1+\varepsilon)^p \|x - u^*\|^p + \frac{3^p}{\varepsilon^{p-1}} \cdot \frac{1}{w(I_x \cap \widetilde{X})} \sum_{x' \in I_x \cap \widetilde{X}} w(x') \, \|x' - u^*\|^p \\ &\leq (1+\varepsilon)^p \|x - u^*\|^p + \frac{3^p}{\varepsilon^{p-1}} \cdot \frac{1}{\alpha'' w(X)} \sum_{x' \in I_x \cap \widetilde{X}} w(x') \, \|x' - u^*\|^p\end{aligned}$$

where $\alpha'' = (1 - \alpha - \alpha')$. Therefore,

$$\begin{aligned}\sum_{x \notin \widetilde{X}} w(x) \|\varphi(x) - v^*\|^p &\leq (1+\varepsilon)^p \sum_{x \notin \widetilde{X}} w(x) \|x - u^*\|^p \\ &+ \frac{3^p}{\varepsilon^{p-1}} \cdot \frac{1}{\alpha'' w(X)} \sum_{\substack{x \notin \widetilde{X} \\ x' \in I_x \cap \widetilde{X}}} w(x) w(x') \|x' - u^*\|^p.\end{aligned}$$

Focusing on the second term, we have

$$\begin{aligned}\frac{1}{\alpha'' w(X)} \sum_{\substack{x \notin \widetilde{X} \\ x' \in I_x \cap \widetilde{X}}} w(x) w(x') \|x' - u^*\|^p &\leq \frac{1}{\alpha'' w(X)} \sum_{x \notin \widetilde{X}} w(x) \sum_{x' \in \widetilde{X}} \|x' - u^*\|^p \\ &\leq \frac{1}{\alpha'' w(X)} \sum_{x \notin \widetilde{X}} w(x) \mathrm{cost}_p(X) \\ &\leq \mathrm{cost}_p(X) \, \frac{w(X \setminus \widetilde{X})}{\alpha'' w(X)} \\ &\leq 2\alpha' \, \mathrm{cost}_p(X)\end{aligned}$$

using the fact that $\alpha'' \geq 1/2$. Putting everything together gives us

$$\begin{aligned}\mathrm{cost}_p(Y) &\leq (1+\varepsilon)^p \sum_{x \in X} w(x) \|x - u^*\|^p + \frac{3^p}{\varepsilon^{p-1}} \cdot 2\alpha' \mathrm{cost}_p(X) \\ &\leq (1 + 3^{p+2}\alpha') \mathrm{cost}_p(X)\end{aligned}$$

for ε sufficiently small. □

Finally, with the above lemmas in hand, we are ready to prove Theorem B.1.

Proof of Theorem B.1. Given Lemma B.8, the proof follows identically as the proof of Theorem 3.2 in [MMR19] by applying Lemma B.8 to the maps $(1+\varepsilon)\varphi$ and $(1+\varepsilon)\varphi^{-1}$. □

Next we prove Theorem B.12, which shows that the cost of each "cluster" in the barycenters problem (i.e. the cost of the weighted flow from points in the input distributions to *one* of the support points in the barycenter) is preserved. We will make use of *distortion graphs*, which quantify the level of distortion of pairwise distances resulting from a dimensionality reduction map.

Definition B.9. *Let $\pi : \mathbb{R}^d \to \mathbb{R}^m$ and X be a set of points in $\mathbb{R}^d$. A distortion graph G with vertex set X is a graph where two points $u, v \in X$ are joined by an edge if the distance between u and v is distorted by a factor at least $1 + \varepsilon$ by π.*

We define the following concept of an everywhere sparse graph to be a generalization to weighted graphs of the concept introduced by [MMR19].

Definition B.10. *Let $G = (V, E)$ be a graph with vertex weights given by $w : V \to \mathbb{R}^{\geq 0}$. Let $N(u)$ denote the neighborhood of a vertex u. G is α-everywhere sparse if*

$$\sum_{v \in N(u)} w(v) \leq \alpha w(V) = \alpha \sum_{v \in V} w(v)$$

for all $u \in V$.

Finally, we impose some additional requirements on the dimensionality reduction map. These essentially say that, even when a pair of points are distorted by the reduction map, the distortion is not too large in expectation.

Definition B.11. *For $\varepsilon > 0$, $\delta \in (0,1)$, a random map $\pi : \mathbb{R}^d \to \mathbb{R}^m$ is an (ε, δ)-dimension reduction if*

$$\frac{1}{1+\varepsilon}\|x-y\| \le \|\pi(x)-\pi(y)\| \le (1+\varepsilon)\|x-y\|,$$

with probability at least $1-\delta$ for every $x, y \in \mathbb{R}^d$. For $p \ge 1$, π is an $(\varepsilon, \delta, \alpha)$-dimension reduction if it additionally satisfies

$$\mathbb{E}\left[\mathbb{1}\{\|\pi(x)-\pi(y)\| > (1+\varepsilon)\|x-y\|\}\left(\frac{\|\pi(x)-\pi(y)\|^p}{\|x-y\|^p} - (1+\varepsilon)^p\right)\right] \le \alpha.$$

We say π is a standard dimension reduction if the parameters $(\varepsilon, \delta, \alpha)$ permit $\delta \le \exp(-C\varepsilon^2 d)$ and $\alpha \le \exp(-C\varepsilon^2 d)$ for $d \ge C'p/\varepsilon^2$ for some absolute constants $C, C' > 0$.

In the theorem below, we just consider one point ν^j in the support of the barycenter and corresponding points (and weights) $S_j, w_j(\cdot)$ assigned to ν^j. We show that a random $(\varepsilon, \delta, \alpha)$-standard dimensionality reduction map roughly preserves the cost of the assignment. The proof is similar to Theorem 3.4 in [MMR19], but we instead use a weighted version of the distortion graphs.

Theorem B.12. *Let $\mu_1, \ldots, \mu_k$ be an instance of the Wasserstein barycenter problem with the L_p objective. Define $X = \bigcup_{i=1}^k \operatorname{supp}(\mu_i)$. Let π be a random $(\varepsilon, \delta, \alpha)$-standard dimensionality reduction map and let $\nu_* = \sum_{j=1}^n b_j \delta(c_j^*)$ be the Wasserstein barycenter. Furthermore, let $\mathcal{C}^* = (C_1^*, \ldots, C_n^*)$ denote the solution to the minimum flow problem for the μ_is to ν^* with corresponding weight functions $w_j^*(\cdot)$ (see Definition 2.1).*

Let $\mathcal{C} = (C_1, \ldots, C_n)$ with corresponding weight functions $w_j(\cdot)$ be any solution (possibly random that depends on π) to the Wasserstein barycenter problem in the sense of Definition 2.1. Let $C = C_j$ be any fixed cluster in $\mathcal{C}$, and further suppose that $\alpha \le 1/10^{p+1}$ and $\delta \le \min(\alpha^7/600, \alpha/n)$. Then with probability at least $1 - \eta - \binom{n}{2}\delta$,

$$\begin{aligned}\operatorname{cost}_p(\pi(C)) &\le A(\operatorname{cost}_p(C) + c_{\eta\varepsilon\alpha}\operatorname{cost}_p(\mathcal{C}^*))\\ \operatorname{cost}_p(C) &\le A(\operatorname{cost}_p(\pi(C)) + c_{\eta\varepsilon\alpha}\operatorname{cost}_p(\mathcal{C}^*)),\end{aligned}$$

where $A = (1+\varepsilon)^{3p-2}(1 + 3^{p+2}\alpha^{1/(p+1)})$ and $c_{\eta\varepsilon\alpha} = \frac{5(1+\varepsilon)^p\alpha}{\eta\varepsilon^{p-1}}$.

Proof. We use a similar outline to the proof of Theorem 3.4 in [MMR19]. Fix $C = C_j$ and let $w_C(\cdot) \equiv w_j(\cdot)$. Let $\mathcal{E}$ be the event that all distances between the points c_i^* are preserved within a $(1+\varepsilon)$-approximation, so that $\mathbf{Pr}[\mathcal{E}] \ge 1 - \binom{n}{2}\delta$ under a random JL projection (or other random standard dimensionality reduction projection). We thus condition the remainder of the proof on the event $\mathcal{E}$. Let $C^\circ \subset C$ be the subset of C whose distances to each center c_i^* are preserved within a $(1+\varepsilon)$-approximation, so that

$$C^\circ = \{x \in C : \pi \text{ preserves the distance between } x \text{ and each } c_i^* \text{ within a factor of } (1+\varepsilon)\}.$$

Note that for a particular $x \in X$, we have that $\mathbf{Pr}[x \in C \setminus C^\circ] \le n\delta$, since by a union bound, the probability that the distance between x and some center c_i^* is distorted by more than a $(1+\varepsilon)$-approximation is at most $n\delta$. Let $G[C^\circ]$ be the graph induced by C° on the distortion graph G. By Theorem B.5, there exists a set $C' \subset C^\circ$ such that $G[C']$ is α-everywhere sparse and $\mathbf{Pr}[x \in C^\circ \setminus C'] \le \alpha$. Thus,

$$\mathbf{Pr}[x \in C \setminus C'] \le \mathbf{Pr}[x \in C \setminus C^\circ] + \mathbf{Pr}[x \in C^\circ \setminus C'] \le n\delta + \alpha \le 2\alpha.$$

We define $g(x)$ to be the identity mapping if $x \in C'$ (so that x is a vertex of the α-everywhere sparse graph) and otherwise, we define $g(x)$ to be the weighted multiset of the assignment of x to each point in the support of the optimal Wasserstein barycenter $\mathcal{C}^*$:

$$g(x) = \begin{cases}(w_C(x), x), & \text{if } x \in C'\\ \{(w_C(x)\cdot r_i(x), c_i^*)\}_{i\in[n]}, & \text{if } x \notin C',\end{cases}$$

where $r_i(x)$ is the ratio of the weight of x that is assigned to the point c_i^*. That is, $r_i(x) = w_i^*(x)/a(x)$, where recall $a(x) =$ weight of x in μ_i. Note that since $\sum_{i=1}^n w_i^*(x) = a(x)$ in order for $w_i^*(\cdot)$ to define a valid solution to the min flow problem, we have $\sum_{i=1}^n r_i(x) = 1$.

Let $\tilde{C} = g(C)$ be a multiset, so that every weighted point in $\tilde{C}$ is either assigned to a point in C' or assigned to some point(s) c_i^* in $\mathcal{C}^*$. Let $\tilde{c}$ be the optimal center for $\tilde{C}$. That is,

$$\tilde{c} = \operatorname*{argmin}_c \sum_{(w,y)\in\tilde{C}} w\|y-c\|^p.$$

Observe that since $G[C']$ is α-everywhere sparse, then the map π $(1+\varepsilon)$-approximates the distances from every $x \in C'$ to (1) at least a $(1-\alpha)$ weighted fraction of the points in C' and (2) to all of the points in the barycenter c_i^*. Conditioning on the event $\mathcal{E}$ so that all pairwise distances between the points c_i^* are approximated within a $(1+\varepsilon)$ factor, then by Theorem B.1,

$$\frac{1}{D}\text{cost}_p(\tilde{C}) \le \text{cost}_p(\pi(\tilde{C})) \le D\text{cost}_p(\tilde{C}), \tag{7}$$

where $D = (1+\varepsilon)^p(1+3^{p+2}\alpha^{1/(p+1)})$. Because $\tilde{c}$ is the optimal center for $\tilde{C}$, we further have

$$\text{cost}_p(C) \le \sum_{x\in C} w_C(x)\|x-\tilde{c}\|^p, \qquad \text{cost}_p(\tilde{C}) = \sum_{x\in C}\sum_{y\in g(x)} w\|y-\tilde{c}\|^p,$$

where we denote each ordered pair in $g(x)$ as (w,y). We compare each term in the right hand side of the relationships for $\text{cost}_p(C)$ and $\text{cost}_p(\tilde{C})$. For $x \in C'$, we have that $g(x) = x$ and $w = w_C(x)$ so the contributions of each term in both summations are the same. For $x \notin C'$, the contributions are $w_C(x)\|x-\tilde{c}\|^p$ and $w_C(x)\sum_{i=1}^n r_i(x)\|c_i^*-\tilde{c}\|^p$ respectively. Hence,

$$\text{cost}_p(C)-(1+\varepsilon)^{p-1}\text{cost}_p(\tilde{C}) \le \sum_{x\in C\setminus C'}\left(w_C(x)\|x-\tilde{c}\|^p - (1+\varepsilon)^{p-1}w_C(x)\sum_{i=1}^n r_i(x)\|c_i^*-\tilde{c}\|^p\right).$$

By triangle inequality, we have that $\|x-\tilde{c}\| \le \|x-c_i^*\| + \|c_i^*-\tilde{c}\|$ for each i. By Lemma B.7 with $r=1$,

$$r_i(x)\|x-\tilde{c}\|^p \le (1+\varepsilon)^{p-1}r_i(x)\|c_i^*-\tilde{c}\|^p + \left(\frac{1+\varepsilon}{\varepsilon}\right)^{p-1} r_i(x)\|x-c_i^*\|^p \tag{8}$$

for each i. Recalling that $\sum_{i=1}^n r_i(x) = 1$ and summing over inequality (8), we have

$$\|x-\tilde{c}\|^p \le (1+\varepsilon)^{p-1}\sum_{i=1}^n r_i(x)\|c_i^*-\tilde{c}\|^p + \left(\frac{1+\varepsilon}{\varepsilon}\right)^{p-1}\sum_{i=1}^n r_i(x)\|x-c_i^*\|^p.$$

Thus,

$$\text{cost}_p(C) - (1+\varepsilon)^{p-1}\text{cost}_p(\tilde{C}) \le \left(\frac{1+\varepsilon}{\varepsilon}\right)^{p-1}\sum_{x\in C\setminus C'} w_C(x)\sum_{i=1}^n r_i(x)\|c_i^*-x\|^p.$$

By similar reasoning, we also have

$$\text{cost}_p(\tilde{C}) - (1+\varepsilon)^{p-1}\text{cost}_p(C) \le \left(\frac{1+\varepsilon}{\varepsilon}\right)^{p-1}\sum_{x\in C\setminus C'} w_C(x)\sum_{i=1}^n r_i(x)\|c_i^*-x\|^p \tag{9}$$

$$\text{cost}_p(\pi(C)) - (1+\varepsilon)^{p-1}\text{cost}_p(\pi(\tilde{C})) \le \left(\frac{1+\varepsilon}{\varepsilon}\right)^{p-1}\sum_{x\in C\setminus C'} w_C(x)\sum_{i=1}^n r_i(x)\|\pi(c_i^*)-\pi(x)\|^p \tag{10}$$

$$\text{cost}_p(\pi(\tilde{C})) - (1+\varepsilon)^{p-1}\text{cost}_p(\pi(C)) \le \left(\frac{1+\varepsilon}{\varepsilon}\right)^{p-1}\sum_{x\in C\setminus C'} w_C(x)\sum_{i=1}^n r_i(x)\|\pi(c_i^*)-\pi(x)\|^p. \tag{11}$$

Along with (7), we have that

$$\begin{aligned}
\text{cost}_p(\pi(C)) &= \left[\text{cost}_p(\pi(C)) - (1+\varepsilon)^{p-1}\text{cost}_p(\pi(\tilde{C}))\right] + (1+\varepsilon)^{p-1}\text{cost}_p(\pi(\tilde{C})) \\
&\leq (10) + (1+\varepsilon)^{p-1} D\text{cost}_p(\tilde{C}) \qquad \text{(due to (7))} \\
&= (10) + (1+\varepsilon)^{p-1} D\left[\left(\text{cost}_p(\tilde{C}) - (1+\varepsilon)^{p-1}\text{cost}_p(C)\right) + (1+\varepsilon)^{p-1}\text{cost}_p(C)\right] \\
&\leq (10) + (1+\varepsilon)^{p-1} D\left[(9) + (1+\varepsilon)^{p-1}\text{cost}_p(C)\right].
\end{aligned}$$

Substituting the right-hand sides of (10) and (9) into the above and factoring out $A = (1+\varepsilon)^{2(p-1)}$, if we define

$$R_x = w_C(x)\sum_{i=1}^{n}\left(\underbrace{r_i(x)\|c_i^* - x\|^p}_{\text{(A)}} + \underbrace{r_i(x)\|\pi(c_i^*) - \pi(x)\|^p}_{\text{(B)}}\right) \tag{12}$$

we see that

$$\text{cost}_p(\pi(C)) \leq A\left(\text{cost}_p(C) + \varepsilon^{1-p}\sum_{c\in C\setminus C'} R_x\right). \tag{13}$$

A similar calculation yields

$$\text{cost}_p(C) \leq A\left(\text{cost}_p(\pi(C)) + \varepsilon^{1-p}\sum_{c\in C\setminus C'} R_x\right), \tag{14}$$

and inequalities (13) and (14) simultaneously hold with probability at least $1 - \binom{n}{2}\delta$.

Finally, we prove that $\varepsilon^{1-p}\sum_{x\in C\setminus C'} \mathbb{1}\{\mathcal{E}\}R_x \leq c_{\eta\varepsilon\alpha}\text{cost}_p(\mathcal{C}^*)$ with probability at least $1-\eta$, by first showing that $\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{x\in C\setminus C'} R_x\right] \leq 5(1+\varepsilon)^p\alpha\text{cost}_p(\mathcal{C}^*)$ and then applying Markov's inequality. We will bound the sum of the (B) terms first. Observe that

$$\|\pi(c_i^*) - \pi(x)\|^p \leq (1+\varepsilon)^p\|c_i^* - x\|^p + \max(\|\pi(c_i^*) - \pi(x)\|^p - (1+\varepsilon)^p\|c_i^* - x\|^p, 0).$$

Furthermore, we have

$$\begin{aligned}
&\max(\|\pi(c_i^*) - \pi(x)\|^p - (1+\varepsilon)^p\|c_i^* - x\|^p, 0) = \\
&\qquad \mathbb{1}\{\|\pi(c_i^*) - \pi(x)\| > (1+\varepsilon)\|c_i^* - x\|\}\left(\|\pi(c_i^*) - \pi(x)\|^p - (1+\varepsilon)^p\|c_i^* - x\|^p\right).
\end{aligned}$$

Thus by Definition B.11, we have that for every $x \in C$ and any $i \in [n]$,

$$\mathbb{E}\left[\max(\|\pi(c_i^*) - \pi(x)\|^p - (1+\varepsilon)^p\|c_i^* - x\|^p, 0)\right] \leq \alpha\|c_i^* - x\|^p.$$

Combining these two bounds, we see that

$$\begin{aligned}
\mathbb{E}\left[\sum_{x\in C\setminus C'} w_C(x)\sum_{i=1}^{n} r_i(x)\,\|\pi(c_i^*) - \pi(x)\|^p\right] &\leq (1+\varepsilon)^p\,\mathbb{E}\left[\sum_{x\in C\setminus C'} w_C(x)\sum_{i=1}^{n} r_i(x)\|c_i^* - x\|^p\right] \\
&\quad + \mathbb{E}\left[\sum_{x\in C\setminus C'} w_C(x)\sum_{i=1}^{n} r_i(x)\max(\|\pi(c_i^*) - \pi(x)\|^p - (1+\varepsilon)^p\|c_i^* - x\|^p, 0)\right].
\end{aligned}$$

The second term in the RHS can be bounded by

$$\begin{aligned}
&\mathbb{E}\left[\sum_{x\in C\setminus C'} w_C(x)\sum_{i=1}^{n} r_i(x)\max(\|\pi(c_i^*)-\pi(x)\|^p-(1+\varepsilon)^p\|c_i^*-x\|^p,0)\right]\\
&\le \sum_{x\in X}\sum_{i=1}^{n} w_C(x)r_i(x)\mathbb{E}\left[\max(\|\pi(c_i^*)-\pi(x)\|^p-(1+\varepsilon)^p\|c_i^*-x\|^p,0)\right]\\
&\le \alpha\sum_{x\in X}\sum_{i=1}^{n} w_i^*(x)\|c_i^*-x\|^p\\
&\le \alpha\,\mathrm{cost}_p(\mathcal{C}^*),
\end{aligned}$$

where we have replaced the sum over $C\setminus C'$ by a larger sum over all X and used the fact that $w_C(x)r_i(x)\le a(x)r_i(x)=w_i^*(x)$ for the weight $w_i^*(x)$ that is assigned to the point c_i^* in the actual barycenter. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\sum_{x\in C\setminus C'} w_C(x)\sum_{i=1}^{n} r_i(x)\,\|\pi(c_i^*)-\pi(x)\|^p\right] &\le (1+\varepsilon)^p\,\mathbb{E}\left[\sum_{x\in C\setminus C'} w_C(x)\sum_{i=1}^{n} r_i(x)\,\|c_i^*-x\|^p\right]\\
&\quad+\alpha\,\mathrm{cost}_p(\mathcal{C}^*).
\end{aligned}\tag{15}$$

By linearity of expectation, we also have

$$\begin{aligned}
\mathbb{E}\left[\sum_{x\in C\setminus C'} w_C(x)\|x-c_i^*\|^p\right] &= \mathbb{E}\left[\sum_{x\in X}\mathbb{1}\{x\in C\setminus C'\}w_C(x)\|x-c_i^*\|^p\right]\\
&= \sum_{x\in X}\mathbf{Pr}\left[x\in C\setminus C'\right]\,w_C(x)\|x-c_i^*\|^p\\
&\le 2\alpha\sum_{x\in X} w_C(x)\|x-c_i^*\|^p.
\end{aligned}$$

It then follows that

$$\begin{aligned}
\mathbb{E}\left[\sum_{x\in C\setminus C'} w_C(x)\sum_{i=1}^{n} r_i(x)\|x-c_i^*\|^p\right] &= \sum_{i=1}^{n} r_i(x)\mathbb{E}\left[\sum_{x\in C\setminus C'} w_C(x)\|x-c_i^*\|^p\right]\\
&\le \sum_{i=1}^{n} r_i(x)\cdot 2\alpha\sum_{x\in X} w_C(x)\|x-c_i^*\|^p\\
&\le 2\alpha\sum_{i=1}^{n}\sum_{x\in X} w_i^*(x)\|x-c_i^*\|^p\\
&= 2\alpha\,\mathrm{cost}_p(\mathcal{C}^*)
\end{aligned}\tag{16}$$

where we again use the fact that $w_C(x)r_i(x)\le a(x)r_i(x)=w_i^*(x)$. Notice that the bound obtained by (16) suffices to bound the first term in (15) *and* the (A) terms from (12). Combining bounds (15) and (16) with the definition of R_x in (12), we obtain

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}\{\mathcal{E}\}\sum_{x\in C\setminus C'} R_x\right] &= \mathbb{E}\left[\sum_{x\in C\setminus C'} w_C(x)\left(\sum_{i=1}^{n} r_i(x)\|c_i^*-x\|^p+r_i(x)\|\pi(c_i^*)-\pi(x)\|^p\right)\right]\\
&\le 2\alpha\mathrm{cost}_p(\mathcal{C}^*)+(1+\varepsilon)^p\cdot 2\alpha\mathrm{cost}_p(\mathcal{C}^*)+\alpha\mathrm{cost}_p(\mathcal{C}^*)\\
&\le 5(1+\varepsilon)^p\alpha\,\mathrm{cost}_p(\mathcal{C}^*).
\end{aligned}$$

By Markov's inequality, we have that

$$\mathbf{Pr}\left[\varepsilon^{1-p}\mathbb{1}\{\mathcal{E}\}\sum_{x\in C\setminus C'}R_x \geq c_{\eta\varepsilon\alpha}\text{cost}_p(\mathcal{C}^*)\right] \leq \varepsilon^{1-p}\cdot\frac{5(1+\varepsilon)^p\alpha\text{cost}_p(\mathcal{C}^*)}{c_{\eta\varepsilon\alpha}\text{cost}_p(\mathcal{C}^*)} = \eta.$$

Recalling that bounds (13) and (14) hold with probability at least $1-\binom{n}{2}\delta$, by a union bound we have

$$\begin{aligned}\text{cost}_p(\pi(C)) &\leq A(\text{cost}_p(C)+c_{\eta\varepsilon\alpha}\text{cost}_p(\mathcal{C}^*))\\ \text{cost}_p(C) &\leq A(\text{cost}_p(\pi(C))+c_{\eta\varepsilon\alpha}\text{cost}_p(\mathcal{C}^*)),\end{aligned}$$

with probability at least $1-\eta-\binom{n}{2}\delta$. □

We now show that the cost of any valid solution to the Wasserstein barycenter problem (again in the sense of Definition 2.1) is roughly preserved under a random $(\varepsilon,\delta,\alpha)$-standard dimensionality reduction map. The proof is similar to Theorem 3.5 in [MMR19], but we again use a weighted version of the distortion graphs.

Theorem B.13. *Let $\mu_1,\ldots,\mu_k$ be k discrete distributions with support size n on $\mathbb{R}^d$, and let $X=\bigcup_{i=1}^k \text{supp}(\mu_i)$. Given $\varepsilon\in(0,1/4)$ and $\delta\in(0,1)$ and $\pi:\mathbb{R}^d\to\mathbb{R}^m$ with*

$$m = O\left(\frac{\log\frac{n}{\delta}+p\log\frac{1}{\varepsilon}+p^2}{\varepsilon^2}\right),$$

then with probability at least $1-\delta$, we have that simultaneously for every solution $\mathcal{C}=(C_1,\ldots,C_n)$ and corresponding weight functions $w_j(\cdot)$ of X,

$$\begin{aligned}\text{cost}_p(\pi(\mathcal{C})) &\leq (1+\varepsilon)^{3p}\text{cost}_p(\mathcal{C})\\ (1-\varepsilon)\text{cost}_p(\mathcal{C}) &\leq (1+\varepsilon)^{3p-1}\text{cost}_p(\pi(\mathcal{C})).\end{aligned}$$

Proof. We define $\theta = \min(\varepsilon^{p+1}3^{-(p+1)(p+2)}, \delta\varepsilon^p/(10n(1+\varepsilon)^{4p-1}, 1/10^{p+1})$. Let $\delta' \leq \min(\theta^7/600,\theta/n)$, $\binom{n}{2}\delta'\leq\delta/2$, $\alpha\leq\delta$, and $\eta\leq\frac{\delta}{2}$. Let the constant in m be sufficiently large, so that π is a random $(\varepsilon,\delta',\alpha)$-standard dimensionality reduction map. Note that the constant m is independent of the quantities $\varepsilon,\delta',\alpha$ due to the properties of a standard dimensionality reduction map. Then we seek to apply Theorem B.12 and note that with these values of ε, δ', and α, we have $A\leq(1+\varepsilon)^{3p-1}$ and $c_{\eta\varepsilon\alpha}\leq\varepsilon/n$.

Let $\mathcal{E}$ be the event that we have

$$\text{cost}_p(\pi(\mathcal{C})) \leq (1+\varepsilon)^{3p}\text{cost}_p(\mathcal{C}) \tag{17}$$

$$(1-\varepsilon)\text{cost}_p(\mathcal{C}) \leq (1+\varepsilon)^{3p-1}\text{cost}_p(\pi(\mathcal{C})) \tag{18}$$

simultaneously any valid solution $\mathcal{C}$ to the Wasserstein barycenter problem. Suppose that the event $\mathcal{E}$ does not occur, so that there exists a solution $\mathcal{C}=\{C_1,\ldots,C_n\}$ and corresponding weight functions $w_j(\cdot)$ that violates (17) or (18). If (17) fails to hold, then

$$\begin{aligned}\sum_{i=1}^n \text{cost}_p(\pi(C_i)) &= \text{cost}_p(\pi(\mathcal{C}))\\ &\geq (1+\varepsilon)^{3p}\text{cost}_p(\mathcal{C})\\ &\geq ((1+\varepsilon)^{3p-1}+\varepsilon)\text{cost}_p(\mathcal{C})\\ &\geq A\left(\sum_{i=1}^n \text{cost}_p(C_i)\right)+\varepsilon\text{cost}_p(\mathcal{C}).\end{aligned}$$

Similarly, if (18) fails to hold, then

$$\begin{aligned}\sum_{i=1}^n \text{cost}_p(C_i) &= (1-\varepsilon)\text{cost}_p(\mathcal{C})+\varepsilon\text{cost}_p(\mathcal{C})\\ &\geq (1+\varepsilon)^{3p-1}\text{cost}_p(\pi(\mathcal{C}))+\varepsilon\text{cost}_p(\mathcal{C})\\ &\geq A\left(\sum_{i=1}^n \text{cost}_p(\pi(C_i))\right)+\varepsilon\text{cost}_p(\mathcal{C}).\end{aligned}$$

It follows that there exists some $i \in [n]$ such that at least one of the following inequalities holds:

$$\mathrm{cost}_p(\pi(C_i)) \geq A\,\mathrm{cost}_p(C_i) + \frac{\varepsilon}{n}\,\mathrm{cost}_p(\mathcal{C})$$
$$\mathrm{cost}_p(C_i) \geq A\,\mathrm{cost}_p(\pi(C_i)) + \frac{\varepsilon}{n}\,\mathrm{cost}_p(\mathcal{C}).$$

Let $\mathcal{C}^*$ be the optimal solution (i.e. the actual Wasserstein barycenter) for $\mu_1, \ldots, \mu_k$. In particular, this means that $\mathrm{cost}_p(\mathcal{C}^*) \leq \mathrm{cost}_p(\mathcal{C})$. Then one of the following inequalities must hold:

$$\mathrm{cost}_p(\pi(C_i)) \geq A\,\mathrm{cost}_p(C_i) + \frac{\varepsilon}{n}\,\mathrm{cost}_p(\mathcal{C}^*)$$
$$\mathrm{cost}_p(C_i) \geq A\,\mathrm{cost}_p(\pi(C_i)) + \frac{\varepsilon}{n}\,\mathrm{cost}_p(\mathcal{C}^*).$$

By Theorem B.12, one of these inequalities can hold with probability at most $\eta + \binom{n}{2}\delta'$. Since $\eta \leq \frac{\delta}{2}$ and $\binom{n}{2}\delta' \leq \frac{\delta}{2}$, it follows that $\mathcal{E}$ occurs with probability at least $1 - \delta$. □

Finally, we show how to rescale the parameters of Theorem B.13 to prove Theorem 4.1.

Proof of Theorem 4.1. Observe that Theorem B.13 with a rescaling of $\varepsilon' := (1+\varepsilon)^{1/(3p)-1} = O(\varepsilon/p)$ immediately implies the desired claim. □

Fast dimensionality reduction. The dimensionality reduction maps of Theorem B.13 generally require multiplication by a dense matrix of (scaled) subgaussian random variables. Thus for $\delta = O(1)$ and $p = O(1)$, applying the dimensionality reduction map using rectangular matrix multiplication takes $O\left(\frac{dkn\log n}{\varepsilon^2}\right)$ time. We provide a tradeoff between runtime and dimension using the following observation:

Theorem B.14. *[MMR19] There exists a family of $(\varepsilon, \delta, \alpha)$-dimensionality reduction maps $\pi : \mathbb{R}^d \to \mathbb{R}^m$ with $m = O\left(\frac{p^6}{\varepsilon^2}\log^2\frac{k}{\varepsilon}\delta\right)$ with runtime $O(d\log d)$ on an input vector $v \in \mathbb{R}^d$.*

We describe the construction of π in Theorem B.14 as in [AC09, MMR19]. By a standard padding with zeros argument, we first assume that d is a power of two. We define D to be a diagonal $d \times d$ matrix with i.i.d. uniform signs, i.e., ± 1 entries. We define H to be a normalized Hadamard transform so that H is an orthogonal matrix with all entries $\pm\frac{1}{\sqrt{d}}$ and Hv can be computed in $O(d\log d)$ time. We define S to be a diagonal "sampling" matrix with i.i.d. entries, so that $\mathbf{Pr}\left[S_{i,i} = \frac{\sqrt{d}}{m}\right] = \frac{m}{d}$ and $\mathbf{Pr}\left[S_{i,i} = 0\right] = 1 - \frac{m}{d}$. Let $\Pi : \mathbb{R}^d \to \mathbb{R}^d$ be defined by $\Pi = SHD$ and note that the expected dimension of the image of Π is m. Then π is defined to be the image of Π, conditioned on the event that the dimension of the image of Π is at most $10d$ [AC09, MMR19]. Hence, we obtain the following fast dimensionality reduction:

Corollary B.15 (Fast dimensionality reduction)**.** *Let $\mu_1, \ldots, \mu_k$ describe an instance of the Wasserstein barycenter problem with the L_p objective in $\mathbb{R}^d$. Given $\varepsilon \in (0, 1/4)$ and $\delta \in (0, 1)$, there exists $\pi : \mathbb{R}^d \to \mathbb{R}^m$ with*

$$m = O\left(\frac{p^6}{\varepsilon^2}\log^2\frac{k}{\varepsilon}\delta\right),$$

that uses $O(d\log d)$ runtime to apply the mapping to each point and provides the same guarantees as Theorem B.13.

Alternative proof for Theorem 4.1. An anonymous NeurIPS 2021 reviewer pointed out the following alternative proof for Theorem 4.1. Given the k distributions $\mu_1, \ldots, \mu_k$ in $\mathbb{R}^d$, consider each distribution μ_i as a multiset U_i of $\mathbb{R}^d$ of size M for a sufficiently large M. Then the Wasserstein barycenter problem can be rewritten as the optimization problem

$$\min \sum_{i=1}^{k} \frac{1}{M} \sum_{u \in U_i} \|u - c_{f_i(u)}\|_2,$$

for an assignment function $f_i : U_i \to [n]$, subject to the constraints:

1. $c_1, \ldots, c_n \in \mathbb{R}^d$
2. $\frac{1}{M}\sum_{u \in U_i} \mathbf{1}[f_i(u) = j] = a_j$, where $\sum_{j \in [n]} a_j = 1$.

Setting U to be the multi-set defined by the union of all U_i with $i \in [k]$ and $f : U \to [n]$ defined by $f(u) = f_i(u)$ for $u \in U_i$, the above optimization can be further rewritten as

$$\min \frac{1}{M} \sum_{u \in U} \|u - c_{f(u)}\|_2,$$

subject to the same constraints. Since this is a constrained k-median clustering problem and [MMR19] show that the cost of *every* clustering is preserved to within a $(1+\varepsilon)$-factor under a projection to $O\left(\frac{1}{\varepsilon^2}\log n\right)$ dimensions, then the optimal clustering under the above constraints are also preserved to within a $(1+\varepsilon)$-factor.

C Proofs for Section 5

We need the following theorem which relates the size of coresets, obtained from importance sampling according to sensitivity values, to the pseudo-dimension of a related function class. Sensitivity sampling has been used to design coresets for many problems in machine learning such as support vector machine, Gaussian mixture models, projective clustering, principal component analysis, M-estimators, Bayesian logistic regression, and generative adversarial networks, e.g., see recent surveys on coresets such as [BLK17, Fel20].

From Theorem 5.5, we now need to bound the following two things to obtain a coreset.

1. The total sensitivity S,
2. The pseudo-dimension of $\mathcal{F}$.

(Note that Theorem 5.2 in [BLK17, Lan18] is stated for coresets of general functions, we specialize it to the case of Wasserstein Barycenters.)

We begin by bounding (1), the total sensitivity S. To do so, we need to define a function s that informs how we sample the distributions in M. The following lemma shows that it suffices to use a constant factor approximation to the best barycenter solution to perform the sampling.

Lemma C.1. *Let $\alpha, p \geq 1$ and let ν' be an α-approximate solution to the p-Wasserstein Barycenter problem for the set M of distributions in $\mathbb{R}^d$ with support size at most n. That is,*

$$\frac{1}{|M|}\sum_{\mu \in M} W(\mu, \nu')^p \leq \alpha \frac{1}{|M|}\sum_{\mu \in M} W(\mu, \nu^*)^p$$

where ν^* is the optimal barycenter distribution. Then the sensitivity $\sigma(\mu)$ for $\mu \in M$ defined as in Definition 5.3 is bounded by

$$\sigma(\mu) \leq s(\mu) \leq \frac{\alpha 2^{p-1} W(\mu, \nu')^p}{\frac{1}{|M|}\sum_{\widetilde{\mu} \in M} W(\widetilde{\mu}, \nu')^p} + \alpha 4^{p-1} + 4^{p-1}.$$

Furthermore, it holds that

$$\mathfrak{S} \leq \alpha(4^{p-1} + 2^{p-1}) + 4^{p-1}.$$

Remark C.2. *Note that p is typically $O(1)$; for example $p = 1$ or $p = 2$ are the most common choices.*

Now our goal is to bound (2), the pseudo-dimension of the function class $\mathcal{F}$ in Theorem 5.5. First we relate pseudo-dimension to VC dimension of related threshold functions and then we state a result relating the VC dimension to the algorithmic complexity of computing these threshold functions.

Lemma C.3 (Pseudo-dimension to VC dimension, Lemma 10 in [LFKF18])**.** *For any $f \in \mathcal{F}$, let B_f be the indicator function of the region on or below the graph of f, i.e., $B_f(x, y) = sgn(f(x) - y)$. The pseudo-dimension of $\mathcal{F}$ is equivalent to the VC-dimension of the subgraph class $B_{\mathcal{F}} = \{B_f \mid f \in \mathcal{F}\}$.*

Then we need the following theorem that relates VC dimension of a function class to its computational complexity.

Lemma C.4 (Theorem 8.14 in [AB99]). *Let $h : \mathbb{R}^a \times \mathbb{R}^b \to \{0, 1\}$, determining the class*

$$\mathcal{H} = \{x \to h(\theta, x) : \theta \in \mathbb{R}^a\}.$$

Suppose that any h can be computed by an algorithm that takes as input the pair $(\theta, x) \in \mathbb{R}^a \times \mathbb{R}^b$ and returns $h(\theta, x)$ after no more than t of the following operations:

- *arithmetic operations $+, -, \times$, and $/$ on real numbers,*
- *jumps conditioned on $>, \geq, <, \leq, =$, and $=$ comparisons of real numbers, and*
- *output $0, 1$,*

then the VC dimension of $\mathcal{H}$ is $O(a^2t^2 + t^2 a \log a)$.

Combining the previous lemmas lets us prove the following theorem. At a high level, we are instantiating Lemma C.4 with the complexity of *computing* any function in the function class $\mathcal{F}$ defined in 5.5. This is a similar proof strategy used in [LFKF18] to control the coreset size of a different problem (coresets for Gaussian mixture models) in the proof of their Theorem 2. However, there is a flaw in their argument as we believe that they incorrectly apply the pseudo-dimension argument to a slightly different function class. We propose a fix in our proof below.

Finally note that the function class $\mathcal{F}$ depends on the the sampling function s. For our purposes, we use the sampling function defined in Lemma C.1 which samples according to a fixed $O(1)$ approximate solution.

Theorem C.5. *Consider the set of functions $\mathcal{F}$ defined in Theorem 5.5 for $s : M \to \mathbb{R}$ as defined in Lemma C.1. The pseudo-dimension of $\mathcal{F}$ is $O(n^8 d^2)$.*

Altogether, we can prove the following bound on the size of coresets for the Wasserstein Barycenter problem which states that k, the number of distributions, can be reduced to $\mathrm{poly}(n, d)$ for constant p.

Theorem C.6. *Let $\delta, \varepsilon \in (0, 1)$. Let ν' be an α approximation to the p-Wasserstein Barycenter problem for a set M of distributions in $\mathbb{R}^d$ with support size at most n and let $s : M \to \mathbb{R}$ be defined as in Lemma C.1. Consider sampling a subset $K \subseteq M$ of size $\widetilde{\Omega}(\alpha 4^{p-1} n^8 d^4/\varepsilon^2)$ where $\widetilde{\Omega}$ hides logarithmic factors. Then K satisfies Definition 5.1 with probability $1 - \delta$.*

Proof. The result follows from instantiating Theorem 5.5 with the bound of S from C.1 and the pseudo-dimension bound in C.5. □

Remark C.7. *Again we remark that we are not optimizing for the exact constants in the exponents in Theorem C.6. There are several places where such optimizations can possibly be made. For example, using a faster algorithm than the Hungarian algorithm to argue about the pseudo-dimension bound in Theorem C.5. However, any such optimizations would result in coresets of size poly(n, d) if we are to use the sensitivity sampling framework.*

Proof of Lemma C.1. Let ν denote an arbitrary barycenter distribution. For any $\mu \in M$, the triangle inequality gives us

$$W(\mu, \nu)^p \leq 2^{p-1}(W(\mu, \nu')^p + W(\nu', \nu)^p)$$

where we have used the fact that $(x + y)^p \leq 2^{p-1}(x^p + y^p)$ for non-negative x, y and $p \geq 1$. Using a similar reasoning, we have

$$W(\nu', \nu)^p \leq 2^{p-1}(W(\nu', \mu') + W(\mu', \nu))$$

for every $\mu' \in M$. Averaging over all μ' gives us

$$W(\nu', \nu)^p \leq \frac{2^{p-1}}{|M|} \sum_{\mu' \in M} (W(\nu', \mu')^p + W(\mu', \nu)^p).$$

It follows that

$$\frac{W(\mu,\nu)^p}{\frac{1}{|M|}\sum_{\widetilde{\mu}\in M} W(\widetilde{\mu},\nu)^p} \leq \frac{2^{p-1}W(\mu,\nu')^p}{\frac{1}{|M|}\sum_{\widetilde{\mu}\in M} W(\widetilde{\mu},\nu)^p} + \frac{\frac{4^{p-1}}{|M|}\sum_{\mu'\in M}\left(W(\nu',\mu')^p + W(\mu',\nu)^p\right)}{\frac{1}{|M|}\sum_{\widetilde{\mu}\in M} W(\widetilde{\mu},\nu)^p}$$

$$\leq \frac{\alpha 2^{p-1}\, W(\mu,\nu')^p}{\frac{1}{|M|}\sum_{\widetilde{\mu}\in M} W(\widetilde{\mu},\nu')^p} + \frac{\frac{\alpha 4^{p-1}}{|M|}\sum_{\mu'\in M} W(\nu',\mu')^p}{\frac{1}{|M|}\sum_{\widetilde{\mu}\in M} W(\widetilde{\mu},\nu')^p} + 4^{p-1}$$

$$\leq \frac{\alpha 2^{p-1}\, W(\mu,\nu')^p}{\frac{1}{|M|}\sum_{\widetilde{\mu}\in M} W(\widetilde{\mu},\nu')^p} + \alpha 4^{p-1} + 4^{p-1} =: s(\mu)$$

where we have used the fact that ν' is an α-approximation to the optimal barycenter and thus,

$$\sum_{\widetilde{\mu}\in M} W(\widetilde{\mu},\nu)^p \geq \sum_{\widetilde{\mu}\in M} W(\widetilde{\mu},\nu^*)^p \geq \frac{1}{\alpha}\sum_{\widetilde{\mu}\in M} W(\widetilde{\mu},\nu')^p$$

by assumption on ν^* and ν'. This gives us

$$\mathfrak{S} \leq \frac{1}{|M|}\sum_{\mu\in M} s(\mu) \leq 4^{p-1} + \alpha 4^{p-1} + \frac{\alpha 2^{p-1}}{|M|}\,\frac{\sum_{\mu\in M} W(\mu,\nu')^p}{\frac{1}{|M|}\sum_{\widetilde{\mu}\in M} W(\widetilde{\mu},\nu')^p} = \alpha(4^{p-1}+2^{p-1}) + 4^{p-1}.$$

Since μ and ν were arbitrary, the result follows. □

Proof of Theorem C.5. Let $\nu \in N$ where N is the set of all possible barycenter distributions with support size n, as defined in Theorem 5.5 and let $r \in \mathbb{R}$. Let M be a set of k different distributions on $\mathbb{R}^d$, each with support size at most n. Then for $x \in M$, we define $h : N \times \mathbb{R} \times M \to \{0,1\}$ by $h(\nu, r, x) = h_{\nu,r}(x) = \mathbb{1}\{W(x,\nu)^p/s(x) \geq r\}$.

We remark that the conceptually similar proof of Theorem 2 in [LFKF18] used to bound the coreset sizes of Gaussian mixture models erroneously omits the function $s(x)$ in the definition of h above.

Now let the corresponding function class $\mathcal{H}$ be defined as

$$\mathcal{H} = \{h_{\nu,r} : M \to \{0,1\} \mid \nu \in N, r \in \mathbb{R}\}.$$

Note that computing $W(x,\nu)^p$ is equivalent to computing the minimum cost bipartite matching between the weighted points of x and ν with edge costs coming from the Euclidean metric raised to the pth power. By the well known Hungarian algorithm, this can be computed in $O(n^3 + n^2 d)$ arithmetic steps where the first term is from the Hungarian algorithm and the second term is to compute the edge costs between x and ν. Furthermore, computing $s(x)$ can also be done in $O(n^3 + n^2 d)$ since we need to find the cost of the matching between x and ν' where ν' is the approximate solution used to define s in Lemma C.1 (the other terms of $s(x)$ are constant).

Therefore by Lemma C.4, the VC dimension of $\mathcal{H}$ is at most $O((nd)^2 \cdot (n^3 + n^2 d)^2) = O(n^8 d^4)$ since we need $O(nd)$ variables to define x and ν. Now note that the function class $\mathcal{H}$ is equivalent to the function class $\mathcal{J}$ defined as

$$\mathcal{J} = \{f_{\nu,r} : M \to \{0,1\} \mid \nu \in N, r \in \mathbb{R}\}$$

where

$$f_{\nu,r} = \mathbb{1}\left\{\frac{W(\cdot,\nu)^p}{s(\cdot)\sum_{m\in M} W(m,\nu)^p} \geq r\right\}$$

This is because we are letting r range over all all the reals in the definition of $\mathcal{H}$. Therefore it also follows that the VC dimension of class $\mathcal{J}$ is $O(n^8 d^4)$. Finally by Lemma C.3, the pseudo-dimension of $\mathcal{F}$ as defined in Theorem 5.5 can be bounded by $O(n^8 d^4)$. □

D Lower Bound Proofs for Section 4.1

We now turn to proving lower bounds, showing that our dimensionality reduction is optimal up to constant factors. To begin, we first need two auxiliary results regarding random linear transformations.

The proof of Theorem 9 in [KMN11] states the following:

Theorem D.1. *[KMN11] Let $M : \mathbb{R}^d \to \mathbb{R}^m$ be a linear transformation with $d > 2m$ and $\varepsilon > 0$ sufficiently small. Then for a randomly chosen unit vector $u \in \mathbb{R}^d$*

$$\mathbf{Pr}\left[\|Mu\|^2 < 1 - \varepsilon\right] \geq \exp(-O(m\varepsilon^2 + 1))).$$

By applying rotational invariance of a standard Gaussian, we arrive at the following corollary of Theorem D.1.

Corollary D.2. *Let $M : \mathbb{R}^d \to \mathbb{R}^m$ be a random matrix with i.i.d. entries from $\mathcal{N}(0, \sigma)$, $d > 2m$ and $\varepsilon > 0$ sufficiently small. Then for any vector $u \in \mathbb{R}^d$,*

$$\mathbf{Pr}\left[\|Mu\|^2 < 1 - \varepsilon\right] \geq \exp(-O(m\varepsilon^2 + 1))).$$

We are now equipped to prove Theorem 4.2, restated here for convenience.

Theorem 4.2. *Consider the setup of Theorem 4.1. Any Gaussian matrix used as a dimension reduction map that allows a $(1 + \varepsilon)$-approximation to the optimal Wasserstein barycenter requires dimension $\Omega(\log n/\varepsilon^2)$.*

Proof. Let $t > 0$ be a parameter and $N > 0$ be a sufficiently large constant. Consider the points $S = p_1, p_2, \ldots, p_t, q_1, q_2, \ldots, q_t$ so that $p_i = Ne_i$ and $q_i = (N + 1)e_i$ for each $i \in [t - 1]$, where e_i is the i-th elementary vector. Let $p_t = Ne_t$ and $q_t = (N + 1 - C\varepsilon)e_t$ for a parameter $C > 0$. Hence we have:

1. $\|p_i - q_i\| = 1$ for each $i \in [t - 1]$.
2. $\|p_t - q_t\| = 1 - C\varepsilon$.
3. $\min_{i \neq j}(\|p_i - p_j\|, \|p_i - q_j\|, \|q_i - q_j\|) \geq N(1 - C\varepsilon)\sqrt{2}$.

Let $n = 2t - 1$ and consider the $k = 2t$ distributions $\mu_1, \ldots, \mu_{2t}$ so that for each $i \in [t]$, μ_i has weight $\frac{1}{2t-1}$ on each of the points in S except p_i, at which it has weight zero. Similarly, for each $i \in [t]$, suppose μ_{i+t} has weight $\frac{1}{2t-1}$ on each of the points in S except q_i, at which it has weight zero. Thus, the total weight across all distributions at each of the $2t$ points in S is exactly 1. It can easily be shown that the barycenter of support size at most n has cost $(1 - C\varepsilon)^p$, by choosing the points p_i and q_i for each $i \in [t - 1]$ and then either p_t or q_t.

We now show that for a Gaussian matrix M with dimension $m = \left(\frac{\log n}{1000\varepsilon^2}\right)$, with high probability there exists some $j \in [t-1]$ such that $\|Mp_j - Mq_j\| \leq \|Mp_t - Mq_t\|$. First note that $\|Mp_t - Mq_t\|$ equals $(1 - C\varepsilon)$ times a random variable that follows a Chi-squared distribution with m degrees of freedom. By standard concentration inequalities on the sum of m independent χ^2 variables, e.g., Equation 2.21 in [Wai19], we have that

$$\mathbf{Pr}\left[\chi^2_m \leq (1 - \varepsilon)m\right] \leq \exp\left(-\frac{m\varepsilon^2}{8}\right) \leq 0.01.$$

Thus the probability that $\|Mp_t - Mq_t\| \leq (1 - C\varepsilon)(1 - \varepsilon)$ is at most 0.01.

Moreover we have that for $i \neq [t]$, $p_i - q_i$ is a unit vector in $\mathbb{R}^d$. Thus by Corollary D.2, we have that

$$\mathbf{Pr}\left[\|Mp_i - Mq_i\|^2 < 1 - (4C)\varepsilon\right] \geq \exp(-O(m(4C)\varepsilon^2 + 1)) \geq \frac{1}{n^{1/5}},$$

for $m = \left(\frac{\log n}{1000C^2\varepsilon^2}\right)$. Therefore since $t = \Omega(n)$, we have that there exists $i \neq [t - 1]$ with $\|Mp_i - Mq_i\|^2 < 1 - (4C)\varepsilon$ with probability at least 0.99.

Hence with probability at least 0.98, the optimal clustering in the projected space will be the projection of the points $p_1, \ldots, p_{j-1}, p_{j+1}, \ldots, p_t$, the points $q_1, \ldots, q_{j-1}, q_{j+1}, \ldots, q_t$, and either the point p_j or q_j. Thus in this case, the corresponding cost in the original space is exactly 1, so that the dimension reduction map does not allow a $(1 + \varepsilon)$-approximation to the optimal clustering. □

Next, we turn to proving lower bounds on dimensionality reduction for the optimal transport problem. Again, we first need an auxiliary concentration result for high-dimensional Gaussians.

Lemma D.3. *Let $C \geq 1$ and fix some point v of norm at most C in $\mathbb{R}^d$. Then, if $x \sim \frac{1}{\sqrt{d}} \cdot \mathcal{N}(0, I_d)$ is a d-dimensional scaled multivariate Normal, then $\Pr(\|x - v\| \leq \frac{1}{C}) \geq n^{-1/10}$, if $d \leq \log n/(10C^2)$ and n is sufficiently large.*

Proof of Lemma D.3. By the rotational symmetry of the multivariate normal, assume $v = (r, 0, \ldots, 0)$, where $0 \leq r \leq C$. Then, if $x = (x_1, y)$ for $x_1 \in \mathbb{R}, y \in \mathbb{R}^{d-1}$, then if $r - \frac{1}{2C} \leq x_1 \leq r$ and $\|y\| \leq \frac{1}{2C}$, then we indeed have $\|x - v\| \leq \frac{1}{C}$. Since $\sqrt{d}x_1 \sim \mathcal{N}(0, 1)$ and $r \leq C$, the probability that $r - \frac{1}{2C} \leq x_1 \leq r$ equals the probability that $\mathcal{N}(0, 1) \in [(r - 1/2C)\sqrt{d}, r\sqrt{d}]$, which is at least $\frac{\sqrt{d}}{2C} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-C^2 d/2}$. Moreover, the probability that $\|y\| \leq \frac{1}{2C}$ is at least $\left(\frac{1}{2eC}\right)^d$. Therefore,

$$\begin{aligned}\Pr\left(\|x - v\| \leq \frac{1}{C}\right) &\geq \frac{\sqrt{d}}{2C} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{-C^2 d/2} \cdot \left(\frac{1}{2eC}\right)^d \\ &\geq n^{-1/10},\end{aligned}$$

where the last inequality is true because $d \leq \log n/(10C^2)$ and that n is sufficiently large. □

Proof of Theorem 4.4. Our point sets will be $A \cup B = \{e_i\} \cup \{e_i/2\}$ with the property that e_i and $e_i/2$ will be in different sets and we will alternate the i's such that $e_i \in A$. The optimal matching in $\mathbb{R}^d$ is to match each e_i to $e_i/2$ leading to cost $d/2$.

Our strategy is to show that if we project $A \cup B$ to $m = o(\log d)$ dimensions, then we can find a matching of cost $o(d)$. Towards that end, let $C = \sqrt{\log d/10m} = \omega(1)$ and let π be a random projection to $o(\log d)$ dimensions. First, we will show that points πe_i with $\|\pi e_i\| \leq C$ will have 'many' other points πe_j sufficiently near by so that we can match πe_i to πe_j (assuming they are in different sets). We then show that the points with $\|\pi e_i\| \geq C$ can be disregarded.

More formally, by Lemma D.3, the number of other points e_j such that $\|\pi e_i - \pi e_j\| \leq 1/C$ and e_j is in a different set than e_i is a binomial random variable $B(d - 1, q)$ where $q \geq d^{-1/10}/2$. Therefore the number of such j's is at least $d^{c'}$ for some constant $c' > 0$ except with probability at most $\ll 1/d$. By a union bound, we can assume that every i such that $\|\pi e_i\| \leq C$ has at least $d^{c'}$ other πe_j's such that $\|\pi e_i - \pi e_j\| \leq 1/C$ and e_j is in a different set than e_i. Now consider the following greedy matching procedure to match the points e_i with $\|\pi e_j \leq C\|$ which may not be optimal: for every such πe_i, we try to match it to any πe_j that is within distance $1/C$ greedily (we also map $\pi e_i/2$ to $\pi e_j/2$). We do this until it is no longer possible. Then, we try to match each πe_i to some πe_j within distance $2/C$ greedily until no longer possible. Then, we just match πe_i to $\pi e_i/2$. Note that every possible match contributes $O(1/C)$ to the matching cost so altogether, this greedy matching has cost at most $O(d/C) = o(d)$.

We now want to show that not many of the πe_i will be leftover that have to be matched to $\pi e_i/2$. Consider maximally covering the set of all such πe_i that have to be matched to $\pi e_i/2$ with disjoint balls of radius $1/C$. First, every such πe_i must be in some ball since other wise, it would have been within radius $2/C$ of some $\pi e_{i'}$ and we would have matched them. Now each ball intersects with at least $d^{c'}$ other points in $A \cup B$ by our calculation in the previous paragraph. Therefore, there can be at most $O(d^{1-c'})$ such balls and hence, the matching cost induced by these points is at most $O(Cd^{1-c'}) = o(d)$ as well.

Now we just have to deal with points e_i that satisfy $\|\pi e_i\| \geq C$. If they are not matched already, we just match them to $\pi e_i/2$. The expected cost incurred by one of these edges in the matching is

$$\begin{aligned}\mathbb{E}\left[\|Ge_i\| \cdot \mathbf{1}_{\|Ge_i\| \geq C}\right] &\leq \sqrt{\mathbb{E}\left[\|Ge_i\|^2\right] \cdot \Pr(\|Ge_i\| \geq C)} \\ &\leq \sqrt{1 \cdot \exp\left(-m \cdot (C-1)^2/8\right)} \\ &\leq \exp\left(-(C-1)^2/16\right) \leq \frac{1}{C},\end{aligned}$$

so the total expected cost from these edges is at most $O(d/C) = o(d)$. Finally by an application of Markov's inequality and a union bound, we have that with probability at least $2/3$, we can find a matching in $\mathbb{R}^m$ with cost at most $o(d)$ and hence, the optimal matching in the projected space has cost at most $o(d)$, as desired. □

If $x \in S^{m-1}$ and π is an appropriately normalized Gaussian dimensionality reduction map, then the following statements hold about the distribution of $\|\pi x\|$ [IN07]:

$$\Pr(|\|\pi x\| - 1| \geq t) \leq \exp(-dt^2/8), \tag{19}$$

$$\Pr(\|\pi x\| \leq 1/t) \leq \left(\frac{3}{t}\right)^d. \tag{20}$$

Finally, we prove Theorem 4.3. Note that Theorem 4.4 states that after we perform a random projection to $o(\log n)$ dimensions, the *cost* (i.e., the actual objective numerical value) of the optimal matching in the projected space will be much smaller than the cost of the optimal matching in the original dimension. This highlights that if we just wish to approximate the cost of the matching, we cannot do better than the standard JL lemma dimension bound. Note that given Theorem 4.3 it is still possible that the optimal matching in the projected dimension is approximately equal to the optimal matching in the original dimension since Theorem 4.3 is only addressing the cost. We show in the proof of Theorem 4.3 that this is not the case; the optimal matching in the projected space will induce a poor matching in the original dimension if we project to much fewer than $\log n$ dimensions.

Proof of Theorem 4.3. Many details of this proof follow similarly as in the proof of Theorem 4.4. Let $C = \sqrt{\log d/10m} = \omega(1)$. Our point sets will be $A \cup B = \{e_i \cdot k/C\}$ for all $1 \leq i \leq d$ and $1 \leq k \leq C$. We refer to A and B as "classes" and assume that C is an even integer. The partition of the points is as follows. For a fixed i, the points $e_i \cdot k/C$ will alternate which set they belong to, i.e, e_i/C will be in A, $2e_i/C$ will be in B etc. We will also impose the condition that half of the e_i's will be in A and the other half will be in B. Now note that the optimal matching in $\mathbb{R}^d$ is to just match each $e_i \cdot k/C$ to $e_i \cdot (k+1)/C$ for $1 \leq k \leq C-1$ which results in matching cost $O(d)$.

Now consider a random projection π to $m = o(\log d)$ dimensions. Our strategy is to show that the optimal matching in $\mathbb{R}^m$ will contain many edges between different e_i's which will induce a large matching cost in $\mathbb{R}^d$.

Towards that end, define 'level k' to be the set of points of the form $e_i \cdot k/C$ for some i. First note that if π is a Gaussian random projection, we have that $\|\pi e_i\| \in [1/10, 100]$ with probability at least $1 - \exp(-m/10) - (3/100)^m > 0.06$ from equations (19) and (20). Thus we can say by a standard Chernoff bound that a $\Theta(1)$ fraction of e_i will satisfy $\|\pi e_i\| = \Theta(1)$ with exponentially small failure probability. By Lemma D.3, for each such πe_i , there exists some e_j such that $\|\pi e_i - \pi e_j\| \leq 1/(100C)$ (again up to some exponentially small failure probability). Since the basis vectors are equally partitioned into the two classes, we can further assume that e_j is in a different class than e_i.

Let I be the set of i's such that $\|\pi e_i\| = \Theta(1)$ and there is some j such that $\|\pi e_i - \pi e_j\| \leq 1/(100C)$ and j is in a different class. For each $i \in I$, the distance between $\pi e_i \cdot k/C$ and $\pi e_i \cdot \ell/C$ for any $\ell \neq k$ is at least $\frac{1}{10C}$ but the distance between $\pi e_i \cdot k/C$ and $\pi e_j \cdot k/C$ is at most $\frac{1}{100C}$. Thus at all levels, we can potentially switch the matching between $\pi e_i \cdot k/C$ and $\pi e_i \cdot (k+1)/C$ or $\pi e_i \cdot (k-1)/C$ (if it exists) to $\pi e_i \cdot k/C$ and $\pi e_j \cdot k/C$ and the same for the point that $\pi e_i \cdot k/C$ was matched to. Therefore, almost all except possibly 1 of the indices in I across all levels will be matched to a point that comes from a different basis vector. Thus the pullback cost is at least some absolute constant times

$$\sum_{i \in I} \sum_{k=1}^{C} \frac{k}{C} \geq \frac{C}{2} \cdot |I|,$$

which is least $\Omega(C \cdot d) = \Omega(C \cdot M) = \omega(M)$, as desired. □

E Connections to Constrained Low-Rank Approximation

[CEM+15] previously showed that the problem of k-means clustering can be formulated as a problem of constrained low-rank approximation, a class of problems which also includes the singular value decomposition (SVD). In this section, we show that the problem of computing a Wasserstein barycenter can be also formulated as a problem of constrained low-rank approximation. Thus efficient subroutines that improve the performance of low-rank approximation can also be used to improve the performance of computing a Wasserstein barycenter.

Recall that for an input matrix $\mathbf{A} \in \mathbb{R}^{a \times b}$ and any set S of rank c orthogonal projection matrices in $\mathbb{R}^{a \times a}$, the goal of constrained low-rank approximation is to find

$$\mathbf{P}^* = \operatorname*{argmin}_{\mathbf{P} \in S} \|\mathbf{A} - \mathbf{P}\mathbf{A}\|_F^2.$$

Proof of Theorem 6.1. For each point x, let $w_i(x)$ be the weight of x in distribution w_i and for each $j \in [n]$, let $w_{i,j}(x)$ be the weight of x in distribution w_i that is assigned to barycenter j, so that we have $\sum_{j \in [n]} w_{i,j}(x) = w_i(x)$ and $\sum_x w_i(x) = 1$ for all i. Thus we have the Wasserstein barycenter objective as minimizing

$$\sum_{j \in [n]} \sum_{i \in [k]} w_{i,j}(x) \|x - C_j\|^p.$$

Rewriting the points of μ_i as $x_{i,1}, x_{i,2}, \ldots, x_{i,n}$, then the Wasserstein barycenter objective for $p = 2$ is

$$\min \sum_{j \in [n]} \sum_{i \in [k]} w_{i,j}(x_{i,j}) \|x_{i,j} - C_j\|^2.$$

Thus we can refold the points $x_{i,j}$ into a matrix of size $\mathbf{A} \in \mathbb{R}^{nk \times d}$ so that the first row of $\mathbf{A}$ consists of the d coordinates of $x_{1,1}$ and more generally row $(i-1)n + j$ of $\mathbf{A}$ consists of the d coordinates of $x_{i,j}$.

Suppose without loss of generality that there exists an integer N such that $w_{i,j}$ is a multiple of $1/N$ for each $i \in [k], j \in [n]$. Let $\mathbf{B} \in \mathbb{R}^{Nk \times d}$ so that each row $(i-1)n + j$ of $\mathbf{A}$ consecutively appears $w_{i,j}(x_{i,j})$ times in $\mathbf{B}$. Thus $\mathbf{B}$ is essentially the matrix whose rows encode each point of each distribution, effectively duplicating each point a number of times equal to its weight in the distribution.

We define a clustering $C = \{C_1, \ldots, C_n\}$ so that there exist weights $w_1, \ldots, w_n$ with $\sum_{j \in [n]} w_j = 1$ with the property that for each $i \in [k]$ and $j \in [n]$, there are exactly $w_j N$ points between rows $(i-1)N+1$ and iN inclusive are assigned to cluster j. Intuitively, this corresponds to each barycenter being assigned weight w_j from each distribution. For each $j \in [n]$, let σ_j be the centroid of all the $w_j Nk$ points assigned to C_j and for each $r \in [Nk]$, let $C(r) \in [n]$ be the cluster to which row r is assigned.

Given a clustering $C = \{C_1, \ldots, C_n\}$, we define the cluster indicator matrix $\mathbf{X}_C \in \mathbb{R}^{Nk \times n}$ to be matrix such that row $(i-1)n + j$ in $\mathbf{X}_C$ has entry $\frac{1}{\sqrt{|C_\ell|}}$ in column $\ell \in [n]$ if and only if the corresponding $\frac{1}{N}$ weight of $x_{i,j}$ is assigned to cluster C_ℓ (and entry zero otherwise). Thus there exist weights $w_1, \ldots, w_n$ with $\sum_{j \in [n]} w_j = 1$ such that for each $i \in [k]$ and $j \in [n]$, column j has exactly $w_j N$ nonzero entries between rows $(i-1)N + 1$ and iN inclusive. Note in this interpretation, we further have $|C_\ell| = w_j N$.

Since the columns of $\mathbf{X}_C$ have disjoint support, then the corresponding vectors are orthonormal. Thus $\mathbf{X}_C \mathbf{X}_C^\top$ is a rank n projection matrix and we can write the problem of Wasserstein barycenter as the constrained low-rank approximation

$$\min_C \frac{1}{N} \|\mathbf{A} - \mathbf{X}_C \mathbf{X}_C^\top \mathbf{A}\|_F^2 = \sum_{r \in [Nk]} \|\mathbf{B}_r - \sigma_{C(r)}\|^2.$$

Note that the cluster indicator matrix $\mathbf{X}_C$ is constrained to the set of valid clusters C consistent with assignments of the support points in the Wasserstein barycenter to each distribution. □

F NP Hardness of Approximation of Wasserstein Barycenters

In this section, we show the NP-hardness of finding a Wasserstein barycenter with cost within a multiplicative 1.0013 factor of the cost induced by an optimal Wasserstein barycenter. We first the following statement about the hardness of approximation for k-means clustering:

Theorem F.1. *[LSW17] It is NP-hard to approximate k-means clustering within a multiplicative factor of* 1.0013.

The proof of Theorem F.1 relies on a reduction from the Vertex Cover problem on 4-regular graphs. Namely, [CC06] showed that it is NP-hard to distinguish whether a 4-regular graph G with n vertices has vertex cover size at least $A_{\max}n$ or vertex cover at most $A_{\min}n$, for some absolute constants $A_{\min} < A_{\max}$. [LSW17] transformed a 4-regular graph G into a graph G' and embedded G' into $\mathbb{R}^{3n}$ so that for the optimal k-means clustering cost of G' (where k is a function of n) is at least $C_{\max}$ if the smallest vertex cover of G has size at least $A_{\max}$ and at least $C_{\min}$ if the smallest vertex cover of G has size at most $A_{\min}$. As it turns out, $C_{\max}/C_{\min} = 1.0013$, which shows the NP-hardness of approximating the optimal k-means clustering cost within a factor of 1.0013.

Given a set G' of N points in $\mathbb{R}^{3n}$, let μ be a uniform distribution on the N points in G' such that each point $p \in G'$ has weight $\frac{1}{N}$. Suppose we restrict the barycenter to have support k, where k is the number of centers in the above k-means clustering instance. Then a set of k centers $c_1, \ldots, c_k$ inducing clusters $C_1, \ldots, C_k$ on G' that achieves cost C for k-means clustering on G' translates to a barycenter of support size k that induces optimal transport cost $\frac{C}{N}$, where the weight of c_i in the barycenter is $\frac{|C_i|}{N}$, for each $i \in [k]$.

Thus the optimal k-means clustering on G' has cost C if and only if the Wasserstein barycenter has cost $\frac{C}{N}$. Therefore, we immediately have the proof of Theorem 6.2.