Proceedings of Machine Learning Research vol 99:1–29, 2019 32nd Annual Conference on Learning Theory

# Sample-Optimal Low-Rank Approximation of Distance Matrices

**Piotr Indyk** INDYK@MIT.EDU
*Massachusetts Institute of Technology*

**Ali Vakilian** VAKILIAN@MIT.EDU
*Massachusetts Institute of Technology*

**Tal Wagner** TALW@MIT.EDU
*Massachusetts Institute of Technology*

**David P. Woodruff** DWOODRUF@CS.CMU.EDU
*Carnegie Mellon University*

**Editors:** Alina Beygelzimer and Daniel Hsu

## Abstract

A distance matrix $A \in \mathbb{R}^{n\times m}$ represents all pairwise distances, $A_{ij} = \mathrm{d}(x_i, y_j)$, between two point sets $x_1, ..., x_n$ and $y_1, ..., y_m$ in an arbitrary metric space $(\mathcal{Z}, \mathrm{d})$. Such matrices arise in various computational contexts such as learning image manifolds, handwriting recognition, and multi-dimensional unfolding.

In this work we study algorithms for low-rank approximation of distance matrices. Recent work by Bakshi and Woodruff (NeurIPS 2018) showed it is possible to compute a rank-$k$ approximation of a distance matrix in time $O((n+m)^{1+\gamma}) \cdot \mathrm{poly}\,(k, 1/\epsilon)$, where $\epsilon > 0$ is an error parameter and $\gamma > 0$ is an arbitrarily small constant. Notably, their bound is sublinear in the matrix size, which is unachievable for general matrices.

We present an algorithm that is both simpler and more efficient. It reads only $O((n+m)k/\epsilon)$ entries of the input matrix, and has a running time of $O(n+m) \cdot \mathrm{poly}\,(k, 1/\epsilon)$. We complement the sample complexity of our algorithm with a matching lower bound on the number of entries that must be read by any algorithm. We provide experimental results to validate the approximation quality and running time of our algorithm.

**Keywords:** Low-rank Approximation, Distance Matrix

## 1. Introduction

Computing low-rank approximations of matrices is a classic computational problem, with a remarkable number of applications in science and engineering. Given an $n\times m$ matrix $A$, and a parameter $k$, the goal is to compute a rank-$k$ matrix $A'$ that minimizes the approximation loss $\|A - A'\|_F$. Such an approximation can be found by computing the singular value decomposition of $A$. However, since the input matrix $A$ is often very large, faster approximate algorithms for computing low-rank approximations have been studied extensively (see the surveys (Mahoney, 2011; Woodruff, 2014) and references therein). In particular, it is known for that for several interesting special classes of matrices, one can find an approximate low-rank solution using only a *sublinear* amount of time or samples from the input matrix. This includes algorithms for *incoherent matrices* (Candès and Recht, 2009), *positive semidefinite matrices* (Musco and Woodruff, 2017) and *distance matrices* (Bakshi and Woodruff, 2018). Note that sub-linear time or sampling bounds are not achievable for general

matrices, as a single large entry in a matrix can significantly influence the output, and finding such an entry could take $\Omega(nm)$ time.

In this paper we focus on computing low-rank approximations of distance matrices, i.e., matrices whose entries are induced by distances between points in some metric space. Formally:

**Definition 1.1 (distance matrix)** *A matrix $A \in \mathbb{R}^{n\times m}$ is called a* distance matrix *if there is an associated metric space $(\mathcal{Z}, \mathrm{d})$ with $\mathcal{X} = \{x_1, \ldots, x_n\} \subset \mathcal{Z}$ and $\mathcal{Y} = \{y_1, \ldots, y_m\} \subset \mathcal{Z}$, such that $A_{ij} = \mathrm{d}(x_i, y_j)$ for every $i, j$.*

Distance matrices occur in many applications, such as learning image manifolds (Weinberger and Saul, 2006), image understanding (Tenenbaum et al., 2000), protein structure analysis (Holm and Sander, 1993), and more. The recent survey (Dokmanic et al., 2015) provides a comprehensive list. Common software packages such as Julia, MATLAB or R include operations specifically design to produce or process such matrices.

Motivated by these applications, a recent paper by (Bakshi and Woodruff, 2018) introduced the first sub-linear time approximation algorithm for distance matrices. Their algorithm computes a rank-$k$ approximation of a distance matrix in time $O((n+m)^{1+\gamma}) \cdot \mathrm{poly}\,(k, 1/\epsilon)$, where $\epsilon > 0$ is an error parameter and $\gamma > 0$ is an arbitrarily small constant. Specifically, it outputs matrices $V \in \mathbb{R}^{n\times k}$ and $U \in \mathbb{R}^{k\times m}$, that satisfy an additive approximation guarantee of the form:

$$\|A - VU\|_F^2 \leq \|A - A_k\|_F^2 + \epsilon\|A\|_F^2,$$

where $A_k$ is the optimal low-rank approximation of $A$. The result of (Bakshi and Woodruff, 2018) raises the question whether even faster algorithms for low-rank approximations of distance matrices are possible. This is the problem we address in this paper.

## 1.1. Our results

In this paper we present an algorithm for low-rank approximation of distance matrices that is both simpler and more efficient than the prior work. Specifically, we show:

**Theorem 1.2 (upper bound)** *There is a randomized algorithm that given a distance matrix $A \in \mathbb{R}^{n\times m}$, reads $O((n+m)k/\epsilon)$ entries of $A$, runs in time $\tilde{O}(n+m) \cdot \mathrm{poly}\,(k, 1/\epsilon)$, and computes matrices $V \in \mathbb{R}^{n\times k}, U \in \mathbb{R}^{k\times m}$ that with probability $0.99$ satisfy*

$$\|A - VU\|_F^2 \leq \|A - A_k\|_F^2 + \epsilon\|A\|_F^2. \tag{1}$$

We complement the sample complexity of our algorithm with a matching lower bound on the number of entries of the input matrix that must be read by any algorithm.

**Theorem 1.3 (lower bound)** *Let $k \leq m \leq n$ and $\epsilon > 0$ be such that $k/\epsilon = O(\min(m, n^{1/3}))$. Any randomized and possibly adaptive algorithm that given a distance matrix $A \in \mathbb{R}^{n\times m}$, computes $V \in \mathbb{R}^{n\times k}, U \in \mathbb{R}^{k\times m}$ that satisfy $\|A - VU\|_F^2 \leq \|A - A_k\|_F^2 + \epsilon\|A\|_F^2$, must read at least $\Omega((n+m)k/\epsilon)$ entries of $A$ in expectation. The lower bound holds even for symmetric distance matrices.*

We include an empirical evaluation of our algorithm on synthetic and real data. The results validate that our approach attains good approximation with faster running time than existing methods.

## 1.2. Our techniques

**Upper bound.** On a high level, our algorithm follows the approach of (Bakshi and Woodruff, 2018). The main idea is to use the result of (Frieze et al., 2004), which shows how to compute a solution satisfying Equation 1 in $\tilde{O}(n+m) \cdot \mathrm{poly}\,(k, 1/\epsilon)$ time, assuming the ability to sample a row (or a column) of the matrix $A$ with the probability at least proportional to its squared-norm.[1] Thus the main challenge is to estimate row and column norms in sub-linear time. Although this cannot be done for general matrices, distance matrices have additional structure (imposed by the triangle inequality), which makes the problem easier. Specifically, estimating column norms in distance matrices corresponds to computing, for all $x \in \mathcal{X}$, the sum of distances (squared) from all points in $\mathcal{Y}$ to $x$. (Bakshi and Woodruff, 2018) gives a sampling algorithm that computes an $n^{0.9}$-approximation of those sums in sub-linear time. Since the approximation is pretty rough, they need to sample many more columns than the original algorithm of (Frieze et al., 2004), and then they apply the algorithm recursively to the sampled columns. The recursive nature of the algorithm makes the procedure and its analysis quite complex.

To avoid this issue, one could design an algorithm that computes a *constant-factor* approximation to row and column norms. In the symmetric case $\mathcal{X} = \mathcal{Y}$, this problem has been already studied in (Indyk, 1999). Specifically, the latter paper developed an approximate comparator, which enables determining whether one row norm is approximately greater than another. Using standard sorting algorithms, one can approximately sort the rows by their norms using roughly $\tilde{O}(n)$ comparisons, where each comparison involves sampling roughly $\tilde{O}(1)$ entries of $A$. Together with fully computing the norms of $\tilde{O}(1)$ landmark rows, this approximate sorting yields sufficiently approximations of all row norms. Unfortunately, this approach does not immediately generalize to the asymmetric case, and exceeds the optimal number of samples by a few logarithmic factors.

Our solution is to estimate row and column norms up to a constant factor with a *one-sided* guarantee. Specifically, we construct estimations that (a) do not *underestimate* the true values and (b) the total sum of the estimations is comparable to the sum of the true values. This is sufficient to support a reduction to (Frieze et al., 2004), while making the estimation procedure much simpler. In the symmetric case, our procedure samples a random point $x^*$, and then estimates the sum of the distances from $x$ to $Y$ as a sum of the distance from $x$ to $x^*$ and the average distance from $x^*$ to $Y$. A simple application of the the triangle inequality shows this estimation provides the desired guarantee. We note that this idea was inspired by the construction of core-sets from (Chen, 2009), although the technical development is quite different (and much simpler).

After executing the algorithm of (Frieze et al., 2004), we still need one more step to compute the solution (as their method reports $U$ but not $V$). This amounts to a regression problem that can be solved by standard techniques. To obtain a tight sampling bound that avoids any logarithmic factor in $k$, we use a recent solver of (Chen and Price, 2017).

**Lower bound.** First let us note that an $\Omega(nk)$ lower bound can be easily obtained. It is not hard to see that any $n \times k$ matrix with entries in $\{1, 2\}$ is an (asymmetric) distance matrix, since the triangle inequality is satisfied trivially. If we choose a uniformly random matrix from $\{1, 2\}^{n \times k}$, then any algorithm that computes a matrix satisfying Equation (1) with $\epsilon = \Omega(1)$, must match a $1 - \Omega(\epsilon)$ fraction of the entries exactly, yielding the lower bound.

1. Formally, the probability of selecting a given row $A_i$ should be $\Omega(\|A_i\|^2/\|A\|_F^2)$.

Our $\Omega(nk/\epsilon)$ lower bound is considerably more involved, and uses tools from communication complexity and random matrix theory. For simplicity, let us describe our techniques in the case $k = 1$. Consider the problem of reporting the majority bit of a random binary string of length $r = \Theta(1/\epsilon)$. This requires reading $\Omega(r)$ of the input bits. If we stack together $n$ instances into an $n \times r$ random binary matrix $A$, then reporting the majority bit for a large fraction of rows requires reading $\Omega(nr)$ input bits. This is our target lower bound.

The reduction proceeds by first shifting the values in $A$ from $\{0, 1\}$ to $\{1, 2\}$, so that it becomes an (asymmetric) distance matrix. A naïve rank-1 approximation would be to replace each entry with 1.5, yielding a total squared Frobenius error of $\frac{1}{4}nr$. However, the optimal rank-1 approximation is (essentially) to replace each row by its true mean value instead of 1.5. By anti-concentration of the binomial distribution, in most rows the majority bit appears $\Omega(\sqrt{r})$ times more often than the minority bit. A simple calculation shows this leads to a constant additive advantage per row, and $\Omega(n)$ advantage over the whole matrix, of the optimal rank-1 approximation over the naïve one. Since $\epsilon\|A\|_F^2 = O(n)$, any algorithm that satisfies Equation (1) must attain a similar advantage.

By spectral properties of random matrices, the optimal rank-1 approximation of $A$ is essentially unique. In particular, the largest singular value of $A$ is much larger than the second-largest one. For technical reasons we need to sharpen this separation even further. We accomplish this by augmenting the matrix with an extra row with very large values, which corresponds to augmenting the metric space with an extra very far point. As a result, any algorithm that satisfies Equation (1) must approximately recover the mean values for a large fraction of the rows. This allows us to solve the majority problem, by reporting whether each row mean in the rank-1 approximation matrix is smaller or larger than 1.5. This yields the desired lower bound for asymmetric distance matrices. The result for symmetric distance matrices, and for general values of $k$, builds on similar techniques.

## 2. Preliminaries

Consider a distance matrix $A \in \mathbb{R}^{n\times m}$ induced by two point sets, $x_1, ..., x_n$ and $y_1, ..., y_m$, as defined in Theorem 1.1. If $n = m$ and $x_i = y_i$ for every $i \in [n]$,[2] then we call $A$ a *symmetric* distance matrix. Otherwise we call it a *bipartite* distance matrix. Throughout, $A_k$ denotes the optimal rank-$k$ approximation of $A$.

As mentioned earlier, our algorithm uses two sub-linear time algorithms as subroutines. They are formalized in the following two theorems. The first reduces low-rank approximation to sampling proportionally to row (or column) norms. We use $A_{i,*}$ to denote the $i$th row of $A$.

**Theorem 2.1 (Frieze et al. (2004))** *Let $A \in \mathbb{R}^{n\times m}$ be any matrix. Let $S$ be a sample of $O(k/\epsilon)$ rows according to a probability distribution $(p_1, \ldots, p_n)$ that satisfies $p_i \geq \Omega(1) \cdot \|A_{i,*}\|_2^2/\|A\|_F^2$ for every $i = 1, \ldots, n$. Then, in time $O(mk/\epsilon + \text{poly}\,(k, 1/\epsilon))$ we can compute from $S$ a matrix $U \in \mathbb{R}^{k\times m}$, that with probability 0.99 satisfies*

$$\|A - AU^TU\|_F^2 \leq \|A - A_k\|_F^2 + \epsilon\|A\|_F^2. \tag{2}$$

The second result approximately solves a regression problem while reading only a small number of columns of the input matrix.

2. Throughout we use $[\ell]$ to denote $\{1, \ldots, \ell\}$ for an integer $\ell$.

**Input:** Distance matrix $A \in \mathbb{R}^{n\times m}$. **Output:** Matrices $V \in \mathbb{R}^{n\times k}$ and $U \in \mathbb{R}^{k\times m}$.

1: Choose $i^* \in [n]$ and $j^* \in [m]$ uniformly at random.
2: For each $i = 1, \ldots, n$: $p_i \leftarrow A_{i,j^*}^2 + A_{i^*,j^*}^2 + \frac{1}{m}\sum_{j=1}^m A_{i^*,j}^2$.
3: Sample $O(k/\epsilon)$ rows of $A$ according to the distribution proportional to $(p_1, \ldots, p_n)$.
4: Compute $U$ from the sample, using Theorem 2.1.
5: Compute $V$ from $A$ and $U$, using Theorem 2.2.
6: Return $V, U$.

**Algorithm 1:** Low-rank approximation for distance matrices

**Theorem 2.2 (Chen and Price (2017))** *There is a randomized algorithm that given matrices $A \in \mathbb{R}^{n\times m}$ and $U \in \mathbb{R}^{k\times m}$, reads only $O(k/\epsilon)$ columns of $A$, runs in time $O(mk) + \mathrm{poly}\,(k, 1/\epsilon)$, and returns $V \in \mathbb{R}^{n\times k}$ that with probability $0.99$ satisfies*

$$\|A - VU\|_F^2 \le (1+\epsilon) \min_{X\in\mathbb{R}^{n\times k}} \|A - XU\|_F^2. \tag{3}$$

Since our sampling procedure evaluates the sum of squared distances (rather than just distances), we need the following approximate version of the triangle inequality.

**Claim 2.3** *For every $x, y, z \in \mathcal{Z}$ in a metric space $(\mathcal{Z}, \mathrm{d})$, $\mathrm{d}(x,y)^2 \le 2(\mathrm{d}(x,z)^2 + \mathrm{d}(z,y)^2)$.*

**Proof** By the triangle inequality, $\mathrm{d}(x,y)^2 \le (\mathrm{d}(x,z) + \mathrm{d}(z,y))^2 = \mathrm{d}(x,z)^2 + 2\mathrm{d}(x,z)\mathrm{d}(z,y) + \mathrm{d}(z,y)^2$. By the inequality of means, $\mathrm{d}(x,z)\mathrm{d}(z,y) \le \frac{1}{2}(\mathrm{d}(x,z)^2 + \mathrm{d}(z,y)^2)$. ■

## 3. Algorithm

In this section we prove Theorem 1.2. The algorithm is stated in Algorithm 1. The main step in the analysis is to provide guarantees for the sampling probabilities $p_i$ computed in Steps 1 and 2 of the algorithm. They are specified by the following theorem.

**Theorem 3.1** *There is a randomized algorithm that given a distance matrix $A \in \mathbb{R}^{n\times m}$, runs in time $O(m+n)$, reads $O(m+n)$ entries of $A$, and outputs sampling probabilities $(p_1, \ldots, p_n)$, that with probability $1-\delta$ satisfy $p_i \ge \Omega(\delta) \cdot \|A_{i,*}\|_2^2/\|A\|_F^2$ for every $i = 1, \ldots, n$.*

**Proof** Let $(\mathcal{Z}, \mathrm{d})$ be the metric space associated with $A$. Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ and $\mathcal{Y} = \{y_1, \ldots, y_m\}$ be the pointsets associated with its rows and its columns, respectively. Choose a uniformly random $i^* \in [n]$ and a uniformly random $j^* \in [m]$. For every $i \in [n]$, the output sampling probabilities are given by

$$p_i = \mathrm{d}(x_i, y_{j^*})^2 + \mathrm{d}(x_{i^*}, y_{j^*})^2 + \frac{1}{m}\sum_{j=1}^m \mathrm{d}(x_{i^*}, y_j)^2.$$

All $p_i$'s can be computed in time $O(n+m)$ and by reading $n+m$ entries of $A$, since they only involve distances between $x_{i^*}$ to $\mathcal{Y}$ and between $y_{j^*}$ to $\mathcal{X}$. For every $i \in [n]$,

$$\begin{aligned}
\|A_{i,*}\|_2^2 = \sum_{j=1}^{m} \mathrm{d}(x_i, y_j)^2 &\le 2\sum_{j=1}^{m} \left(\mathrm{d}(x_i, y_{j^*})^2 + \mathrm{d}(y_{j^*}, y_j)^2\right) && \text{by Claim 2.3} \\
&\le 2\sum_{j=1}^{m} \left(\mathrm{d}(x_i, y_{j^*})^2 + 2\mathrm{d}(x_{i^*}, y_{j^*})^2 + 2\mathrm{d}(x_{i^*}, y_j)^2\right) && \text{by Claim 2.3} \\
&= 2m \cdot \mathrm{d}(x_i, y_{j^*})^2 + 4m \cdot \mathrm{d}(x_{i^*}, y_{j^*})^2 + 4\sum_{j=1}^{m} \mathrm{d}(x_{i^*}, y_j)^2 \\
&\le 4m \cdot p_i.
\end{aligned}$$

On the other hand, in expectation over $i^*$ and $j^*$ we have $\mathbb{E}\left[\mathrm{d}(x_i, y_{j^*})^2\right] = \frac{1}{m}\sum_{j=1}^{m} \mathrm{d}(x_i, y_j)^2$, and $\mathbb{E}\left[\mathrm{d}(x_{i^*}, y_{j^*})^2\right] = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{d}(x_i, y_j)^2$, and $\mathbb{E}\left[\mathrm{d}(x_{i^*}, y_j)^2\right] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{d}(x_i, y_j)^2$. Thus,

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{n} p_i\right] &= \sum_{i=1}^{n}\left(\mathbb{E}\left[\mathrm{d}(x_i, y_{j^*})^2\right] + \mathbb{E}\left[\mathrm{d}(x_{i^*}, y_{j^*})^2\right] + \mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m} \mathrm{d}(x_{i^*}, y_j)^2\right]\right) \\
&= 3n \cdot \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{d}(x_i, y_j)^2 = \frac{3}{m}\|A\|_F^2.
\end{aligned}$$

By Markov's inequality, $\sum_{i=1}^{n} p_i \le \frac{3}{\delta m}\|A\|_F^2$ with probability $1-\delta$. Normalizing the $p_i$'s by their sum yields the theorem. ∎

We remark that if $A$ is a symmetric distance matrix, i.e., $\mathcal{X} = \mathcal{Y}$, the sampling probabilities can be simplified to choosing a single $i^* \in [n]$ uniformly at random, and letting $p_i = \mathrm{d}(x_i, x_{i^*})^2 + \frac{1}{n}\sum_{j=1}^{m} \mathrm{d}(x_{i^*}, x_j)^2$. The proof is similar to the above.

**Proof of Theorem 1.2.** Consider Algorithm 1. By Theorem 3.1, the probabilities computed in Steps 1–2 are suitable for invoking Theorem 2.1. This ensures that the matrix $U$ computed in Steps 3–4 satisfies Equation (2). Theorem 2.2 guarantees that the matrix $V$ computed in Step 5 satisfies Equation (3). Putting these together, we have

$$\begin{aligned}
\|A - VU\|_F^2 &\le (1+\epsilon)\min_{X \in \mathbb{R}^{n\times k}} \|A - X^T U\|_F^2 && \text{by Equation (3)} \\
&\le (1+\epsilon)\|A - AU^T U\|_F^2 \\
&\le (1+\epsilon)\left(\|A - A_k\|_F^2 + \epsilon\|A\|_F^2\right) && \text{by Equation (2)} \\
&\le \|A - A_k\|_F^2 + \epsilon \cdot (2+\epsilon) \cdot \|A\|_F^2 && \text{since } \|A - A_k\|_F^2 \le \|A\|_F^2,
\end{aligned}$$

and we can scale $\epsilon$ by a constant. This proves Equation (1). For the query complexity bound, observe that Theorem 2.1 reads $O(k/\epsilon)$ rows and Theorem 2.2 reads $O(k/\epsilon)$ columns of the matrix, yielding a total of $O((n+m)k/\epsilon)$ queries. Finally, the running time is the sum of runnings times of Theorems 2.1, 2.2 and 3.1. ∎

# 4. Lower Bound

For a clearer presentation, in this section we prove the lower bound in the special case $k = 1$, for distance matrices that can be asymmetric (called *bipartite* in Definition 1.1). This case encompasses the main ideas. The full proof of Theorem 1.3 appears in the appendix. For concreteness, let us formally state the special case that will be proven in this section.

**Theorem 4.1** *Let $n, r, \epsilon$ be such that $r \leq n$ and $1 > \epsilon \geq \Omega(n^{-1/3})$. Any randomized algorithm that given a distance matrix $A \in \mathbb{R}^{n\times r}$, computes $V \in \mathbb{R}^{n\times k}, U \in \mathbb{R}^{k\times r}$ that with probability $2/3$ satisfy $\|A - VU\|_F^2 \leq \|A - A_1\|_F^2 + \epsilon\|A\|_F^2$, must read at least $\Omega(n/\epsilon)$ entries of $A$ in expectation.*

## 4.1. Hard Distribution over Distance Matrices

By Yao's principle, it suffices to construct a distribution over distance matrices, and prove the sampling lower bound for any deterministic algorithm that operates on inputs from that distribution. We begin by defining a suitable distribution over distance matrices and proving some useful properties.

**Hard problem.** In the majority problem, the goal is to compute the majority bit of an input bit-string. We will show the hardness of low-rank approximation via reduction from solving multiple random instances of the majority problem. The sample-complexity hardness of this problem is well-known, and is stated in the following lemma. The proof is included in the appendix.

**Lemma 4.2** *Let $r, t > 0$ be integers. Any deterministic algorithm that gets a uniformly random matrix $S \in \{0,1\}^{t\times r}$ as input, and outputs $s^* \in \{0,1\}^t$ such that for every $i \in [t]$, $\Pr[s^*(i) =$ majority element of $i$th row of $S] \geq 2/3$, must read in expectation at least $\Omega(rt)$ entries of $S$.*

**The distribution.** Given $n$ and $\epsilon > 0$, let $\beta, C > 0$ be constants that will be chosen later. ($\beta$ will be sufficiently small and $C$ sufficiently large.) Let $r = \beta/\epsilon$, and assume w.l.o.g. this is an integer by letting $\epsilon$ be sufficiently smaller. Note that in Lemma 4.2, we can symbolically replace the majority alphabet $\{0,1\}$ with any alphabet of size 2, and here we will use $\{1,2\}$. Let $S \in \{1,2\}^{n\times r}$ be a uniformly random matrix. Let $s_1, \ldots, s_n$ be its rows. We call each of its rows an *instance* (of the majority problem). Thus $S$ is an instance of the random multi-instance majority problem from Lemma 4.2 (with $t = n$). We begin by establishing some of its probabilistic properties.

Our goal is to solve $S$ via reduction to rank-1 approximation of distance matrices. To obtain a distribution over distance matrices, we first take $S$ and randomly permute its rows to obtain a matrix $A$. The random permutation is denoted by $\pi : [n] \to [n]$.[3] Then, we add an additional $(n+1)$th row to $A$, whose entries are all equal $M = \sqrt{Cn}$. The matrix with the added row is denoted by $\bar{A}$.

**Metric properties.** First we show that $\bar{A}$ is indeed a (bipartite) distance matrix.

**Lemma 4.3** *Every supported $\bar{A}$ is a distance matrix.*

3. The random permutation is for a technical reason and does not change the distribution. Specifically, it is to prevent the algorithm in Lemma 4.2 from focusing on a few fixed instances $\{s_i\}_{i=1}^{n'}$,, $n' \ll n$, and never attempt the rest.

**Proof** Consider a symbolic pointset $X = P \cup Q$ where $P \cap Q = \emptyset$, such that $P = \{p_1, \ldots, p_{n+1}\}$ corresponds to the rows of $\bar{A}$, and $Q = \{q_1, \ldots, q_r\}$ to the columns of $\bar{A}$. Our goal is to define a metric $\mathrm{d}$ on $X$ such that $\mathrm{d}(p_i, q_j) = \bar{A}_{ij}$ for every $i \in [n+1]$ and $j \in [r]$. We need to set the rest of the distances such that $\mathrm{d}$ is indeed a metric – that is, such that $\mathrm{d}$ satisfies the triangle inequality. For every $i, i' \in [n]$ we set $\mathrm{d}(p_i, p_{i'}) = 1$. For every $j, j' \in [r]$ we set $\mathrm{d}(q_j, q_{j'}) = 1$. Finally we need to set the distances from $p_{n+1}$. By construction of $\bar{A}$ we already have $\mathrm{d}(p_{n+1}, q_j) = M$ for every $j \in [r]$. We set all the remaining distances, $\mathrm{d}(p_i, p_{n+1})$ for every $i \in [n]$, to also be $M$.

We need to verify that for all distinct triplets $x, y, z \in X$, $\mathrm{d}(x, y) \leq \mathrm{d}(x, z) + \mathrm{d}(z, y)$. Indeed, all distances are in $\{1, 2, M\}$. If $\mathrm{d}(x, y) \in \{1, 2\}$ then the inequality holds for any setting of $\mathrm{d}(x, z)$ and $\mathrm{d}(z, y)$. Otherwise $\mathrm{d}(x, y) = M$, hence necessarily either $x = p_{n+1}$ or $y = p_{n+1}$, and in both cases $\mathrm{d}(x, z) + \mathrm{d}(z, y) \geq \max\{\mathrm{d}(x, z), \mathrm{d}(z, y)\} \geq \mathrm{d}(p_{n+1}, z) = M = \mathrm{d}(x, y)$ as needed. ■

**Probabilistic properties.** By anti-concentration of the binomial distribution, it is known that in a random length-$r$ bistring, the majority bit is likely appear $\Omega(\sqrt{r})$ times more than the other bit.

**Lemma 4.4 (anti-concentration)** *Let $0 < \delta < 1$. Let $s \in \{1, 2\}^r$ be a uniformly random majority instance. Then, for $\gamma = \Omega(\delta)$, the majority element of $s$ appears in it at least $\frac{1}{2}r + \gamma\sqrt{r}$ times with probability at least $1 - \delta$.*

We call an instance $s$ in $S$ *typical* if its majority element appears in it at least $\frac{1}{2}r + \gamma\sqrt{r}$ times, where $\gamma$ is the constant from Lemma 4.4. Otherwise, we call the instance *atypical*.

Let $\Psi_{\text{typical}}$ denote the event there are at least $0.9n$ typical instances. By Markov's inequality, $\Pr[\Psi_{\text{typical}}] \geq 1 - 10\delta$.

**Spectral properties.** We will also require some facts from random matrix theory about the spectrum of $A$. Let $\mathbf{1}$ denote the all-1's vector in $\mathbb{R}^r$. Let $P_{\mathbf{1}}$ denote the orthogonal projection on the subspace spanned by $\mathbf{1}$. The proofs of the following two lemmas are given in the appendix.

**Lemma 4.5** *Suppose $r^3 = O(n)$. With probability $1 - e^{-\Omega(n/r^{3/2})}$, $\|A - A_1\|_F^2 \geq \frac{1}{4}nr - O(n)$.*

In the next lemma and throughout, $\|X\|_2$ denotes the spectral norm of a matrix $X$.

**Lemma 4.6** *Let $\zeta > 0$. Let $Z$ be a rank-1 matrix such that $\|Z\|_2 \leq \zeta\sqrt{n}$. Then with probability $1 - o(1)$, $\|AP_{\mathbf{1}}^{\perp} - Z\|_F^2 \geq \|AP_{\mathbf{1}}^{\perp}\|_F^2 - \zeta \cdot O(n)$.*

Since $r = \beta/\epsilon$ and in Theorem 4.1 we assume $\epsilon^{-3} = O(n)$, Lemma 4.5 is satisfied with probability $1 - o(1)$. Therefore,

**Corollary 4.7** *Denote by $\Psi_{\text{spectral}}$ the event that the conclusions of both Lemmas 4.5 and 4.6 hold. Then $\Pr[\Psi_{\text{spectral}}] \geq 1 - o(1)$, and therefore $\Pr[\Psi_{\text{typical}} \wedge \Psi_{\text{spectral}}] \geq 1 - 10\delta - o(1)$.*

**Bounds on low-rank approximation** Finally we give upper bounds on the approximation error allowed by Equation (1). For every instance $s_i$, let $\mu_i = \frac{1}{r}\sum_{j=1}^{r} s_{ij}$ denote its mean.

**Lemma 4.8** $\|\bar{A} - \bar{A}_1\|_F^2 + \epsilon\|\bar{A}\|_F^2 \leq \sum_{i=1}^{n} \|s_i - \mu_i\mathbf{1}\|_2^2 + (4 + C)\beta n$.

**Proof** Let $\bar{A}^*$ be the matrix in which each row equals $\mathbf{1}$ times the mean of the corresponding row of $A$. Then $\|\bar{A} - \bar{A}_1\|_F^2 \leq \|\bar{A} - \bar{A}^*\|_F^2 = \sum_{i=1}^{n} \|s_i - \mu_i \mathbf{1}\|_2^2$, where the first inequality is since $\bar{A}^*$ has rank 1 (each of its rows is a multiple of $\mathbf{1}$). Note that the sum ranges only up to $n$ and not $n+1$, since in the $(n+1)$th row all entries are equal (to $M$) and thus it contributes 0 to $\|\bar{A} - \bar{A}^*\|_F^2$. This bounds the first summand in the lemma. To bound the second summand, note that each entry in the first $n$ rows of $\bar{A}$ is at most 2, thus contributing in total $4rn$ to $\|\bar{A}\|_F^2$. The final row contributes $rM^2 = Crn$. Recalling that $\epsilon r = \beta$, we have $\epsilon \|A\|_F^2 \leq (4 + C)\beta n$. ■

**Corollary 4.9** $\|\bar{A} - \bar{A}_1\|_F^2 + \epsilon \|\bar{A}\|_F^2 \leq \frac{1}{4} nr + (4 + C)\beta n$.

**Proof** For every $s_i$, its mean $\mu_i$ minimizes the sum of squared differences from a single value, namely $\|s_i - \mu_i \mathbf{1}\|_2^2 = \min_{\nu \in \mathbb{R}} \|s_i - \nu \mathbf{1}\|_2^2$. In particular, $\|s_i - \mu_i \mathbf{1}\|_2^2 \leq \|s_i - 1.5 \cdot \mathbf{1}\|_2^2$. Furthermore, since $s_i \in \{1, 2\}^r$, we have $\|s_i - 1.5 \cdot \mathbf{1}\|_2^2 = r \cdot (\frac{1}{2})^2 = \frac{1}{4} r$. Hence $\sum_{i=1}^{n} \|s_i - \mu_i \mathbf{1}\|_2^2 \leq \frac{1}{4} nr$, and the corollary follows from Lemma 4.8. ■

## 4.2. Invoking the Algorithm

Suppose we have a deterministic algorithm that given $\bar{A}$, returns $\bar{A}' = \bar{a} b^T$, where $\bar{a} \in \mathbb{R}^{n+1}$ and $b \in \mathbb{R}^r$, such that

$$\|\bar{A} - \bar{a} b^T\|_F^2 \leq \|\bar{A} - \bar{A}_1\|_F^2 + \epsilon \|\bar{A}\|_F^2. \tag{4}$$

Let $a \in \mathbb{R}^n$ denote the restriction of $\bar{a}$ to the first $n$ entries. By scaling (i.e., multiplying $\bar{a}$ by a constant and $b$ by its reciprocal), we can assume w.l.o.g. that $a_{n+1} = M$. Since the $(n+1)$th row of $\bar{A}$ equals $M \cdot \mathbf{1}$, we have

$$\|\bar{A} - \bar{a} b^T\|_F^2 = \|A - a b^T\|_F^2 + \|M \cdot \mathbf{1} - a_{n+1} b\|_2^2 = \|A - a b^T\|_F^2 + M^2 \|\mathbf{1} - b\|_2^2.$$

If we rearrange this, and use Equation (4) and Corollary 4.9 as an upper bound on $\|\bar{A} - \bar{a} b^T\|_F^2$ and Lemma 4.5 as a lower bound on $\|A - a b^T\|_F^2$, we get $M^2 \|\mathbf{1} - b\|_2^2 \leq (4 + C)\beta n + O(n)$. Plugging $M = \sqrt{Cn}$,

$$\|\mathbf{1} - b\|_2^2 \leq \left( \frac{4}{C} + 1 \right) \beta + \frac{O(1)}{C} = \frac{O(1)}{C}. \tag{5}$$

This facts yields the following two lemmas, whose full proofs appear in the appendix.

**Lemma 4.10** *We have* $1 - \eta/\sqrt{r} \leq \|b^T P_{\mathbf{1}}\| / \|\mathbf{1}\| \leq 1 + \eta/\sqrt{r}$*, where* $\eta > 0$ *is a constant that can be made arbitrarily small by choosing* $C > 0$ *sufficiently large.*

**Lemma 4.11** $\|a\| = O(\sqrt{n})$.

## 4.3. Solving Majority

We now show how to use $\bar{A}' = \bar{a} b^T$ to solve the majority instance $S$ of the problem in Lemma 4.2. We condition on the intersection of the events $\Psi_{\text{typical}}$ and $\Psi_{\text{spectral}}$. By Corollary 4.7 it occurs with probability at least $1 - 10\delta - o(1)$.

Let $s_1, \ldots, s_n \in \{1, 2\}^r$ denote the random instances in $S$. Recall that we assigned them to rows of $A$ by a uniformly random permutation $\pi$, that is, the $\pi(i)$th row of $A$ equals $s_i$.

We use $a$ to solve the majority problem as follows. For each $s_i$, if $a_{\pi(i)} \leq 1.5$ then we output that the majority is 1, and otherwise we output that the majority is 2. We say that $a$ *solves* the instance $s_i$ if the output is correct. Due to $\pi$ being random, the probability that $A$ solves any instance $s_i$ is identical. Denote this probability by $p$. We need to show that $p \geq 2/3$.

Assume by contradiction that $p < 2/3$. By Markov's inequality, with probability at least $4/5$ we have at least $n/6$ unsolved instances. Since by $\Psi_{\text{typical}}$ there are only $0.1n$ atypical instances, we have at least $n/15$ unsolved typical instances. Denote by $S'$ the set of unsolved typical instances. Consider such instance $s_i \in S'$. Suppose its majority element is 1. Then, since it is typical,

$$\begin{aligned}\|s_i - \mu_i \mathbf{1}\|_2^2 &\leq \|s_i - (1.5 - \gamma/\sqrt{r})\mathbf{1}\|_2^2 \\ &\leq \left(\tfrac{1}{2}r + \gamma\sqrt{r}\right)\left(\tfrac{1}{2} - \gamma/\sqrt{r}\right)^2 + \left(\tfrac{1}{2}r - \gamma\sqrt{r}\right)\left(\tfrac{1}{2} + \gamma/\sqrt{r}\right)^2 \\ &= \tfrac{1}{4}r - \gamma^2. \end{aligned} \tag{6}$$

On the other hand, since $s_i$ is unsolved then $a_{\pi(i)} \geq 1.5$. Hence by Lemma 4.10, $\frac{\|b^T P_{\mathbf{1}}\|_2}{\|\mathbf{1}\|_2} \cdot a_{\pi(i)} \geq 1.5 - \eta/\sqrt{r}$. Therefore, noting that $b^T P_{\mathbf{1}} = \frac{\|b^T P_{\mathbf{1}}\|_2}{\|\mathbf{1}\|_2} \cdot \mathbf{1}$, we have

$$\begin{aligned}\|s_i - a_{\pi(i)} b^T P_{\mathbf{1}}\|_2^2 &= \|s_i - \frac{\|b^T P_{\mathbf{1}}\|_2}{\|\mathbf{1}\|_2} \cdot a_{\pi(i)}\mathbf{1}\|_2^2 \\ &\geq \|s_i - (1.5 - \tfrac{\eta}{\sqrt{r}}) \cdot \mathbf{1}\|_2^2 \\ &\geq \left(\tfrac{1}{2}r + \gamma\sqrt{r}\right)\left(\tfrac{1}{2} - \eta/\sqrt{r}\right)^2 + \left(\tfrac{1}{2}r - \gamma\sqrt{r}\right)\left(\tfrac{1}{2} + \eta/\sqrt{r}\right)^2 \\ &= \tfrac{1}{4}r + \eta^2 - 2\gamma\eta. \end{aligned} \tag{7}$$

Similar calculations yield the same bounds when the majority element is 2. From Equation (6), together with Equation (4) and Lemma 4.8, we get:

$$\|\bar{A} - \bar{a}b^T\|_F^2 \leq \sum_{i=1}^{n} \|s_i - \mu_i \mathbf{1}\|_2^2 \leq \tfrac{1}{15}n(\tfrac{1}{4}r - \gamma^2) + \sum_{s_i \notin S'} \|s_i - \mu_i \mathbf{1}\|_2^2 + (4 + C)\beta n. \tag{8}$$

On the other hand, by Equation (7),

$$\|A - ab^T P_{\mathbf{1}}\|_F^2 = \sum_{i=1}^{n} \|s_i - a_{\pi(i)} b^T P_{\mathbf{1}}\|_2^2 \geq \tfrac{1}{15}n(\tfrac{1}{4}r + \eta^2 - 2\gamma\eta) + \sum_{s_i \notin S'} \|s_i - \mu_i \mathbf{1}\|_2^2. \tag{9}$$

It remains to relate Equations (8) and (9) to derive a contradiction. By the Pythagorean identity, $\|A - ab^T\|_F^2 = \|AP_{\mathbf{1}} - ab^T P_{\mathbf{1}}\|_F^2 + \|AP_{\mathbf{1}}^{\perp} - ab^T P_{\mathbf{1}}^{\perp}\|_F^2$, and

$$\|A - ab^T P_{\mathbf{1}}\|_F^2 = \|AP_{\mathbf{1}} - ab^T P_{\mathbf{1}}^2\|_F^2 + \|AP_{\mathbf{1}}^{\perp} - ab^T P_{\mathbf{1}} P_{\mathbf{1}}^{\perp}\|_F^2 = \|AP_{\mathbf{1}} - ab^T P_{\mathbf{1}}\|_F^2 + \|AP_{\mathbf{1}}^{\perp}\|_F^2.$$

Together, $\|A - ab^T P_{\mathbf{1}}\|_F^2 = \|A - ab^T\|_F^2 + \left(\|AP_{\mathbf{1}}^{\perp}\|_F^2 - \|AP_{\mathbf{1}}^{\perp} - ab^T P_{\mathbf{1}}^{\perp}\|_F^2\right)$. Let us upper-bound both terms. For the first term, we simply use $\|A - ab^T\|_F^2 \leq \|\bar{A} - \bar{a}b^T\|_F^2$. For the second term, note that $\|b^T P_{\mathbf{1}}^{\perp}\|_2 = \|\mathbf{1}P_{\mathbf{1}}^{\perp} - b^T P_{\mathbf{1}}^{\perp}\|_2 \leq \|\mathbf{1} - b\|_2$. Together with Lemma 4.11, $\|ab^T P_{\mathbf{1}}^{\perp}\|_2 \leq O(\sqrt{n}) \cdot \|1 - b\|_2$. Thus by Lemma 4.6, $\left(\|AP_{\mathbf{1}}^{\perp}\|_F^2 - \|AP_{\mathbf{1}}^{\perp} - ab^T P_{\mathbf{1}}^{\perp}\|_F^2\right) \leq O(n) \cdot \|\mathbf{1} - b\|_2$. By Equation (5), the latter is $O(n)/\sqrt{C}$. Plugging both upper bounds,

$$\|A - ab^T P_{\mathbf{1}}\|_F^2 \leq \|\bar{A} - \bar{a}b^T\|_F^2 + O(n) \cdot C^{-1/2}.$$

This relates Equations (8) and (9), yielding

$$\tfrac{1}{15}(\gamma^2+\eta^2) \le \tfrac{2}{15}\cdot\gamma\eta + (4+C)\beta + O(1)\cdot C^{-1/2}.$$

Since $\gamma$ is fixed, choosing $\beta, \eta$ sufficiently small and $C$ sufficiently large leads to a contradiction.

Thus $p \ge 2/3$, meaning the reduction solves each instance in the majority problem $S$ with probability at least $2/3$. Accounting for the conditioning on $\Psi_{\text{typical}}$ and $\Psi_{\text{spectral}}$, the determined low-rank approximation algorithm from Section 4.2 solves a random instance of Lemma 4.2 with probability at least $2/3 - 10\delta - o(1)$ (the constants can be scaled without changing the lower bound). Hence, it requires reading at least $\Omega(n/\epsilon)$ bits from the matrix, which proves Theorem 4.1.

## 5. Experiments

In this section, we evaluate the empirical performance of Algorithm 1 compared to the existing methods in the literature: conventional SVD, the algorithm of (Bakshi and Woodruff, 2018) (BW), and the input-sparsity time algorithm of (Clarkson and Woodruff, 2017) (IS). For SVD we use numpy's linear algebra package.[4] The experimental setup is analogous to that in (Bakshi and Woodruff, 2018). Specifically, we consider two datasets:

- **Synthetic clustering dataset.** This data set is generated using the `scikit-learn` package. We generate $10,000$ points with 200 features and partition the points into 20 clusters. As observed in our experiments, the dataset is expected to have a good rank-20 approximation.
- **MNIST dataset.** The dataset contains $70,000$ handwritten characters, and each is considered a point. We subsample $10,000$ points.

For each dataset we construct a symmetric distance matrix $A_{i,j} = \mathrm{d}(p_i, p_j)$. We use four distances $\mathrm{d}$: Manhattan ($\ell_1$), Euclidean ($\ell_2$), Chebyshev ($\ell_\infty$) and Canberra[5] ($\ell_c$). Figures 1 and 2 show the approximation error for each distance on each dataset, for varying values of the rank $k$. Note that SVD achieves the optimal approximation error. Table 1 lists the running times for $k = 40$. Figure 3 shows the running time of our algorithm for MNIST subsampled to varying sizes, for $k = 40$.

Table 1: Running times (in seconds) of the compared methods for rank $k = 40$ approximation

| | Synthetic | | | | MNIST | | | |
|---|---|---|---|---|---|---|---|---|
| **Metric** | **SVD** | **IS** | **BW** | **This Work** | **SVD** | **IS** | **BW** | **This Work** |
| $\ell_2$ | 398.77 | 8.95 | 1.70 | 1.17 | 398.50 | 34.32 | 4.17 | 1.23 |
| $\ell_1$ | 410.60 | 8.16 | 1.82 | 1.197 | 560.91 | 39.50 | 3.71 | 1.23 |
| $\ell_\infty$ | 427.90 | 9.18 | 1.63 | 1.16 | 418.01 | 39.33 | 4.00 | 1.14 |
| $\ell_c$ | 452.17 | 8.49 | 1.76 | 1.15 | 390.07 | 38.34 | 3.91 | 1.24 |

4. https://docs.scipy.org/doc/numpy-1.15.1/reference/routines.linalg.html. This performs full SVD. The iterative SVD algorithms built into MATLAB and Python yielded errors larger by a few orders of magnitude than the reported methods, so they are not included.
5. The Canberra distance $\mathrm{d}_c$ between vectors $p, q \in \mathbb{R}^n$ is defined as $\mathrm{d}_c(p,q) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}$.

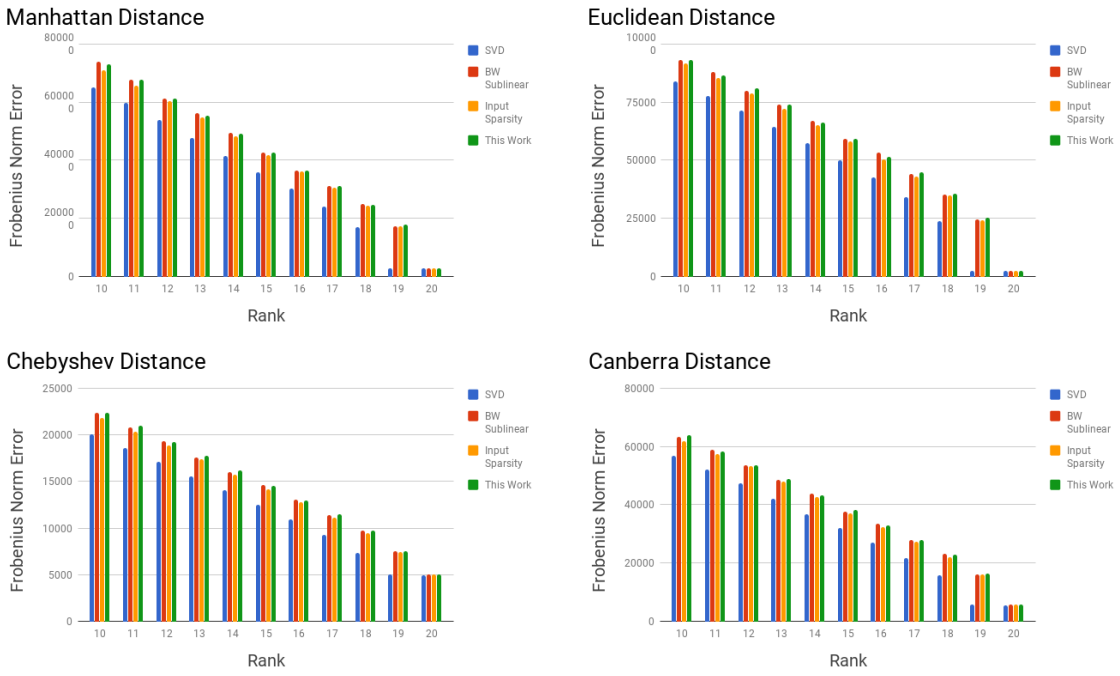


Figure 1: The approximation error of the four algorithms on the synthetic clustering dataset.

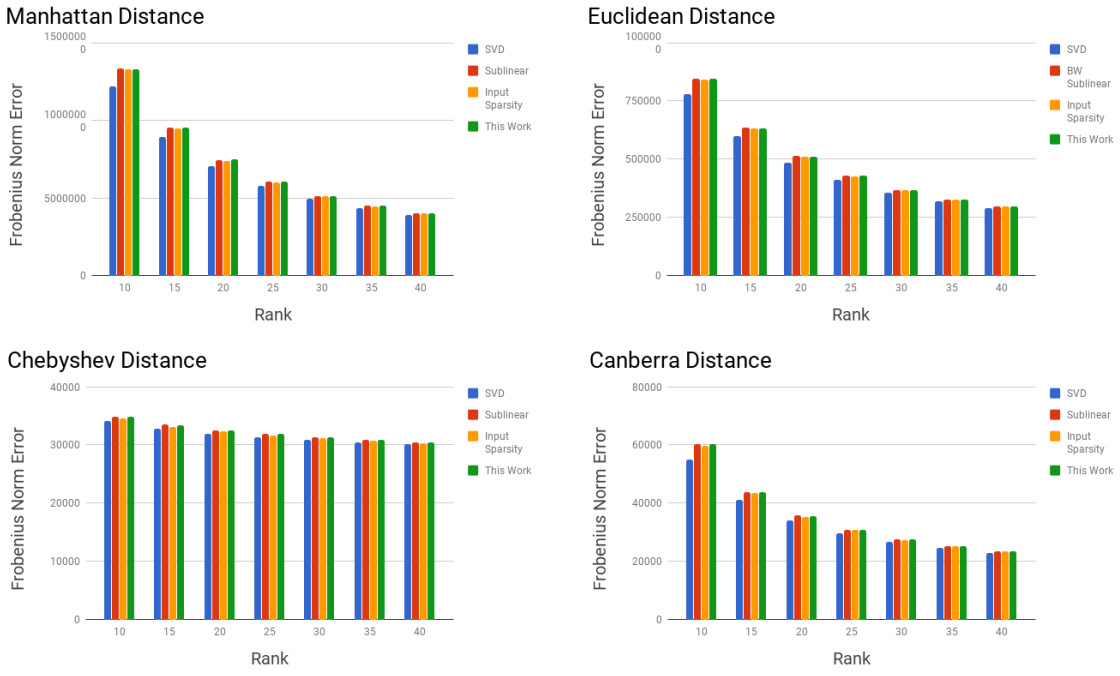


Figure 2: The approximation error of the four algorithms on the MNIST dataset.

## Acknowledgments

P. Indyk, A. Vakilian and T. Wagner were supported by funds from the MIT-IBM Watson AI Lab, NSF, and Simons Foundation. D. Woodruff was supported partly by the National Science Foundation under Grant No. CCF-1815840, and this work was done partly while he was visiting the Simons Institute for the Theory of Computing. The authors would also like to thank Ainesh Bakshi for implementing the algorithm in this paper and producing our empirical results.

## References

Ainesh Bakshi and David Woodruff. Sublinear time low-rank approximation of distance matrices. In *Advances in Neural Information Processing Systems*, pages 3786–3796, 2018.

Mark Braverman and Anup Rao. Information equals amortized communication. *IEEE Transactions on Information Theory*, 60(10):6058–6069, 2014.

Mark Braverman, Ankit Garg, Denis Pankratov, and Omri Weinstein. Information lower bounds via self-reducibility. *Theory of Computing Systems*, 59(2):377–396, 2016.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

Ke Chen. On coresets for k-median and k-means clustering in metric and euclidean spaces and their applications. *SIAM Journal on Computing*, 39(3):923–947, 2009.

Xue Chen and Eric Price. Condition number-free query and active learning of linear families. *arXiv preprint arXiv:1711.10051*, 2017.

Kenneth L Clarkson and David P Woodruff. Low-rank approximation and regression in input sparsity time. *Journal of the ACM (JACM)*, 63(6):54, 2017. (first appeared in STOC'13).

Ivan Dokmanic, Reza Parhizkar, Juri Ranieri, and Martin Vetterli. Euclidean distance matrices: essential theory, algorithms, and applications. *IEEE Signal Processing Magazine*, 32(6):12–30, 2015.

Ohad N Feldheim and Sasha Sodin. A universality result for the smallest eigenvalues of certain sample covariance matrices. *Geometric And Functional Analysis*, 20(1):88–123, 2010.

Alan Frieze, Ravi Kannan, and Santosh Vempala. Fast monte-carlo algorithms for finding low-rank approximations. *Journal of the ACM (JACM)*, 51(6):1025–1041, 2004.

Alan J Hoffman and Helmut W Wielandt. The variation of the spectrum of a normal matrix. In *Selected Papers Of Alan J Hoffman: With Commentary*, pages 118–120. World Scientific, 2003.

Liisa Holm and Chris Sander. Protein structure comparison by alignment of distance matrices. *Journal of molecular biology*, 233(1):123–138, 1993.

Piotr Indyk. A sublinear time approximation scheme for clustering in metric spaces. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 154–159. IEEE, 1999.

Michael W Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

Cameron Musco and David P Woodruff. Sublinear time low-rank approximation of positive semidefinite matrices. In *Foundations of Computer Science (FOCS), 2017 IEEE 58th Annual Symposium on*, pages 672–683. IEEE, 2017.

Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, pages 1576–1602. World Scientific, 2010.

Terence Tao. 254a, notes 3a: Eigenvalues and sums of hermitian matrices. https://terrytao.wordpress.com/2010/01/12/254a-notes-3a-eigenvalues-and-sums-of-hermitian-matrices, 2010.

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

RC Thompson. The behavior of eigenvalues and singular values under perturbations of restricted rank. *Linear Algebra and its Applications*, 13(1-2):69–78, 1976.

Kilian Q Weinberger and Lawrence K Saul. Unsupervised learning of image manifolds by semidefinite programming. *International journal of computer vision*, 70(1):77–90, 2006.

David P Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

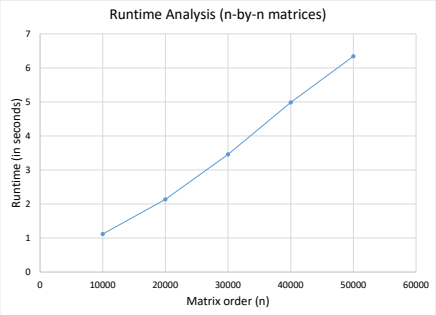


Figure 3: Running time of our algorithm on subsets of MNIST, for $k = 40$.

# Appendix A. Deferred Proofs from Section 4

## A.1. Preliminary Lemmas

The following lemmas are known and we include their proofs for completeness.

**Lemma A.1 (hardness of majority, Lemma 4.2 restated)** *Let $r, t > 0$ be integers. Any deterministic algorithm that gets a uniformly random matrix $S \in \{0,1\}^{t \times r}$ as input, and outputs $s^* \in \{0,1\}^t$ such that for every $i \in [t]$, $\Pr[s^*(i) =$ majority element of $i$th row of $S] \geq 2/3$, must read in expectation at least $\Omega(rt)$ entries of $S$.*

**Proof** The reduction is from the distributional Gap-Hamming communication problem, which is defined as follows. Alice has a bit string $x$ in $\{0,1\}^r$ and Bob has a bit string $y$ in $\{0,1\}^r$, where $x$ and $y$ are independent and uniformly distributed. Let $\Delta(x,y)$ denote their Hamming distance. The goal is to decide whether $\Delta(x,y) \geq \frac{1}{2}r + \sqrt{r}$ or $\Delta(x,y) \leq \frac{1}{2}r - \sqrt{r}$. If neither case holds, then any output is considered successful. The information cost under this distribution is $\Omega(r)$ (Braverman et al., 2016).

Next consider the $r$-fold version of the same problem, i.e., Alice and Bob are given $r$ instances of distributional Gap-Hamming, and they need to solve a constant fraction of them. By a standard direct sum theorem (see e.g. (Braverman and Rao, 2014)), this requires $\Omega(rt)$ bits of communication.

Finally we reduce this problem to the majority problem in lemma statement. Let $X$ denote the xor matrix of Alice's and Bob's matrices. The Gap Hamming problem is equivalent to finding the majority bit over rows of $X$ in which the majority bit appears at least $\frac{1}{2}r + \sqrt{r}$ times (call these rows "typical"). By Lemma 4.4 this happens in a large constant fraction of the rows. Given a black-box algorithm for the majority problem that queries $q$ entries of the input matrix, Alice and Bob can simulate it on $M$ by communicating to each other only those entries of their matrices, which costs them $\Theta(q)$. The algorithm solves a large fraction of the rows, and thus a large fraction of the typical rows. Hence they have solved the Gap Hamming problem, and $q = \Omega(rt)$. ■

**Lemma A.2 (binomial anti-concentration, Lemma 4.4 restated)** *Let $0 < \delta < 1$. Let $s \in \{1,2\}^r$ be a uniformly random majority instance. Then, for $\gamma = \Omega(\delta)$, the majority element of $s$ appears in it at least $\frac{1}{2}r + \gamma\sqrt{r}$ times with probability at least $1 - \delta$.*

**Proof** Let $X \sim \text{Binomial}(r, \frac{1}{2})$. The statement we need to show is equivalent to $\Pr[|X - \frac{1}{2}r| < \frac{1}{2}\gamma\sqrt{r}] < \delta$, or equivalently,

$$\sum_{i=\lfloor \frac{1}{2}r - \frac{1}{2}\gamma\sqrt{r} \rfloor}^{\lceil \frac{1}{2}r + \frac{1}{2}\gamma\sqrt{r} \rceil} \Pr[X = i] < \delta.$$

Let $\gamma = \sqrt{\pi/2} \cdot \delta$. Note that $\Pr[X = i] \leq \Pr[X = \lfloor r/2 \rfloor]$ for every $i$, and $\Pr[X = \lfloor r/2 \rfloor] = 2^{-r}\binom{r}{\lfloor r/2 \rfloor} \leq 1/\sqrt{2\pi r}$ by a known estimate. Therefore, the above left-hand side sum is upper-bounded by $2\gamma\sqrt{r}/\sqrt{2\pi r} = \delta$ as needed. ■

## A.2. Spectral Properties

For the next two lemmas, write $A = 1.5J + B$ where $J$ is the all-1's matrix and $B$ is a matrix with i.i.d. random entries chosen uniformly from $\{-\frac{1}{2}, \frac{1}{2}\}$.

**Lemma A.3 (Lemma 4.5, restated)** *Suppose $r^3 = O(n)$. With probability $1 - e^{-\Omega(n/r^{3/2})}$, $\|A - A_1\|_F^2 \geq \frac{1}{4}nr - O(n)$.*

**Proof** We use a sharp estimate of (Feldheim and Sodin, 2010) on the smallest singular value of $B$ (see also (Rudelson and Vershynin, 2010), eq. (2.5)). It states that with probability $1 - \exp(-\Omega(n/r^{3/2}))$, all $r$ singular values of $B$ are at least $\frac{1}{4}n - O(\sqrt{nr})$. Furthermore, since $A$ is obtained from $B$ by adding a rank-1 matrix $(1.5J)$, then by Theorem 1 of (Thompson, 1976), $A$ has at least $r-1$ singular values which are at least $\frac{1}{4}n - O(\sqrt{nr})$. Therefore $\|A - A_1\|_F^2$, which is the sum of all squared singular values of $A$ except the largest, is at least $(r-2) \cdot (\frac{1}{4}n - O(\sqrt{nr})) = \frac{1}{4}nr - O(n) - O(r^{3/2}\sqrt{n})$. ■

**Lemma A.4 (Lemma 4.6, restated)** *Let $\zeta > 0$. Let $Z$ be a rank-1 matrix such that $\|Z\|_2 \le \zeta\sqrt{n}$. Then with probability $1 - o(1)$, $\|AP_{\mathbf{1}}^{\perp} - Z\|_F^2 \ge \|AP_{\mathbf{1}}^{\perp}\|_F^2 - \zeta \cdot O(n)$.*

**Proof** By the Hoffman-Wielandt inequality (Hoffman and Wielandt, 2003) for singular values[6], any $X, Y \in \mathbb{R}^{n\times r}$ satisfy $\|X - Y\|_F^2 \ge \sum_{j=1}^{r}(\sigma_j(X) - \sigma_j(Y))^2$, where $\sigma_1(X) \ge \ldots \ge \sigma_r(X)$ and $\sigma_1(Y) \ge \ldots \ge \sigma_r(Y)$ are their respective sorted singular values. In particular, letting $X = AP_{\mathbf{1}}^{\perp}$ and $Y = Z$, since $Z$ has rank 1, we have

$$\begin{aligned}\|AP_{\mathbf{1}}^{\perp} - Z\|_F^2 &\ge \sum_{j=2}^{r}(\sigma_j(AP_{\mathbf{1}}^{\perp}))^2 + \Big(\sigma_1(AP_{\mathbf{1}}^{\perp}) - \sigma_1(Z)\Big)^2 \\ &= \|AP_{\mathbf{1}}^{\perp}\|_F^2 - 2\sigma_1(AP_{\mathbf{1}}^{\perp})\sigma_1(Z) + (\sigma_1(Z))^2 \\ &= \|AP_{\mathbf{1}}^{\perp}\|_F^2 - 2\|AP_{\mathbf{1}}^{\perp}\|_2\|Z\|_2 + \|Z\|_2^2.\end{aligned}$$

Therefore it suffices to show that $\|AP_{\mathbf{1}}^{\perp}\|_2 = O(\sqrt{n})$. Indeed, $\|AP_{\mathbf{1}}^{\perp}\|_2 = \|(1.5J + B)P_{\mathbf{1}}^{\perp}\|_2 = \|BP_{\mathbf{1}}^{\perp}\|_2 \le \|B\|_2$, and an upper bound $\|B\|_2 = O(\sqrt{n})$ is well known, e.g., see Proposition 2.4 in (Rudelson and Vershynin, 2010). ■

## A.3. Lemmas from Section 4.2

**Lemma A.5 (Lemma 4.10, restated)** *We have*

$$1 - \frac{\eta}{\sqrt{r}} \le \frac{\|b^T P_{\mathbf{1}}\|}{\|\mathbf{1}\|} \le 1 + \frac{\eta}{\sqrt{r}},$$

*where $\eta > 0$ is a constant that can be made arbitrarily small by choosing $C > 0$ sufficiently large.*

**Proof** By the triangle inequality we have

$$\|\mathbf{1}\|_2 - \|\mathbf{1} - b\|_2 \le \|b\|_2 \le \|\mathbf{1}\|_2 + \|\mathbf{1} - b\|_2.$$

The upper bound implies

$$\|b^T P_{\mathbf{1}}\| \le \|b\| \le \|\mathbf{1}\|_2 + \|\mathbf{1} - b\|_2. \tag{10}$$

The lower bound implies

$$\|\mathbf{1} - b\|_2^2 = \|\mathbf{1}\|_2^2 + \|b\|_2^2 - 2b^T\mathbf{1} \ge \|\mathbf{1}\|_2^2 + \|\mathbf{1}\|_2^2 + \|\mathbf{1} - b\|_2^2 - 2\|\mathbf{1}\|_2\|\mathbf{1} - b\|_2 - 2b^T\mathbf{1},$$

6. See also Exercise 22(v) in (Tao, 2010)

which rearranges to $b^T\mathbf{1} \geq \|\mathbf{1}\|_2^2 - \|\mathbf{1}\|_2\|\mathbf{1} - b\|_2$, implying

$$\|b^T P_{\mathbf{1}}\| = b^T\left(\frac{1}{\|\mathbf{1}\|_2}\mathbf{1}\right) \geq \|\mathbf{1}\|_2 - \|\mathbf{1} - b\|_2. \tag{11}$$

Putting Equations (10) and (11) together,

$$\|\mathbf{1}\|_2 - \|\mathbf{1} - b\|_2 \leq \|b^T P_{\mathbf{1}}\| \leq \|\mathbf{1}\|_2 + \|\mathbf{1} - b\|_2,$$

and the lemma follows since $\|\mathbf{1}\|_2 = \sqrt{r}$ and since by eq. (5), $\|\mathbf{1} - b\|_2$ is a constant that can be made arbitrarily small by choosing $C > 0$ sufficiently large. ■

**Lemma A.6 (Lemma 4.11, restated)** $\|a\| = O(\sqrt{n})$.

**Proof** By the triangle inequality, $\|b\|_2 \geq \|\mathbf{1}\|_2 - \|b - \mathbf{1}\|_2$. Since $\|\mathbf{1}\|_2 = \sqrt{r}$ and $\|b - \mathbf{1}\|_2$ is an arbitrarily small constant by Equation (5), $\|b\|_2 \geq \frac{1}{2}\sqrt{r}$. Thus

$$\|ab^T\|_F^2 = \|a\|_2^2\|b\|_2^2 \geq \tfrac{1}{4}r\|a\|_2^2. \tag{12}$$

We finish by showing that $\|ab^T\|_F^2 = O(nr)$. Indeed,

$$\begin{aligned}
\|A - ab^T\|_F^2 &\leq \|\bar{A} - \bar{a}b^T\|_F^2 \\
&\leq \|\bar{A} - \bar{A}_1\|_F^2 + \epsilon\|\bar{A}\|_F^2 && \text{by Equation (4)} \\
&\leq \tfrac{1}{4}nr + (4 + C)\beta n. && \text{by Corollary 4.9}
\end{aligned}$$

Furthermore $\|A\|_F^2 \leq 4nr$ since each entry of $A$ has absolute value at most 2. Finally, by approximate triangle inequality (Claim 2.3),

$$\|ab^T\|_F^2 \leq 2\|A\|_F^2 + 2\|A - ab^T\|_F^2 = O(nr).$$

With Equation (12) this implies the lemma. ■

## A.4. Lower Bound for Symmetric Distance Matrices

In this section we show that the lower bound in Theorem 4.1 applies also to symmetric distance matrices. The proof is by a reduction to the asymmetric case.

### A.4.1. General $k$

We start by reducing rank-$k$ approximation of asymmetric distance matrices to rank-$(2k + 2)$ approximation of symmetric distance matrices. By tuning $\beta$, suppose w.l.o.g. that $kr = \beta k/\epsilon$ is a divisor of $n$. Let $B \in \mathbb{R}^{n\times kr}$ be an asymmetric distance matrix drawn from the hard distribution. Recall that all entries are in $\{1, 2, 3\}$. We can scale them by half so they are all the interval $[1, 2]$.

We construct a symmetric distance matrix $A \in \mathbb{R}^{2n\times 2n}$. It is partitioned into $n \times n$ blocks,

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

We set its entries as follows. Its main diagonal is all-zeros. $A_{21}$ consists of $n/(kr)$ copies of $B$, concatenated horizontally. $A_{11}$ has all off-diagonal entries set to 1. $A_{12}$ and $A_{22}$ are determined symmetrically. Since all entries are in the interval $[1, 2]$, the triangle inequality is satisfied trivially and thus $A$ is a distance matrix.

We will show here that any rank-$(2k + 2)$ approximation algorithm for $A$ must read at least $\Omega(nk/\epsilon)$ of its entries. Let $\mathbf{0}$ and $\mathbf{1}$ denote the all-0's and all-1's vectors in $\mathbb{R}^n$, respectively. For any $x, y \in \mathbb{R}^n$ let $[x\ y]$ denote their concatenation into a vector in $\mathbb{R}^{2n}$. Let $B_k$ be the optimal rank-$k$ approximation of $B$. Write $B$ as $B = UV^T$ where $U \in \mathbb{R}^{n\times k}$ and $V \in \mathbb{R}^{kr\times k}$. Let $u_1, \ldots, u_k$ be the columns of $U$ and let $v_1, \ldots, v_k$ be the columns of $V^T$. For every $i \in [k]$, let $\bar{v}_i$ be the vector given by concatenating $n/(kr)$ copies of $v_i$. Consider the rank-$(2k + 2)$ approximation of $A$ given by the column vectors $\{[\mathbf{0}\ u_i] : i \in [k]\} \cup \{[\bar{v}_i\ \mathbf{0}] : i \in [k]\} \cup \{[\mathbf{1}\ \mathbf{0}], [\mathbf{0}\ \mathbf{1}]\}$. Let us bound its error in approximating $A$. On $A_{12}$ and $A_{21}$, which contain concatenated copies of $B$, the vectors $\{[\mathbf{0}\ u_i] : i \in [k]\} \cup \{[\bar{v}_i\ \mathbf{0}] : i \in [k]\}$ attain the optimal error. On $A_{11}$ and $A_{11}$, which contain 0's on the diagonal and 1's off the diagonal, the vectors $\{[\mathbf{1}\ \mathbf{0}], [\mathbf{0}\ \mathbf{1}]\}$ attain zero error on the off-diagonal entries, and $2n$ error in total over the diagonal entries. Consequently,

$$\|A - A_{2k+2}\|_F^2 \leq t \cdot \|B - B_1\|_F^2 + 2n,$$

where $t = 2n/(kr)$ is the number of copies of $B$ embedded in $A$.

Suppose we have an algorithm that given $A$, returns a rank-$(2k+2)$ matrix $A'$ that satisfies Equation (1). Let $A'_1, \ldots, A'_t$ denote the restriction of $A'$ to the blocks matching the copies of $B$ embedded in $A$. Thus, $\|A - A'\|_F^2 \geq \sum_{i=1}^t \|B - A'_i\|_F^2$. Furthermore, since $\|A\|_F^2 = \Theta(n^2) = \Theta(tnkr)$ and $\|B\|_F^2 = \Theta(nkr)$, we have $\|A\|_F^2 = O(1) \cdot t \cdot \|B\|_F^2$. Putting everything into Equation (1),

$$\sum_{i=1}^{t} \|B - A'_i\|_F^2 \leq t \cdot \left(\|B - B_k\|_F^2 + O(n) + O(\epsilon) \cdot \|B\|_F^2\right).$$

By averaging, for at least one $i \in [t]$ we have $\|B - A'_i\|_F^2 \leq \|B - B_k\|_F^2 + O(\epsilon) \cdot \|B\|_F^2$. By scaling $\epsilon$ by a constant, $A'$ solves the rank-$k$ approximation problem for $B$. By the proof for the asymmetric case, this requires $\Omega(nk/\epsilon)$ queries to its input.

### A.4.2. $k = 1$

The previous section proves hardness for rank-$k$ approximation of symmetric distance matrices, for $k \geq 4$. For completeness let us also show hardness for the $k = 1$ case, by a somewhat more refined analysis of the reduction in that case.

To this end we slightly modify the construction of $A$ from the previous section. We draw $B \in \mathbb{R}^{n\times r}$ from the hard distribution for asymmetric distance matrices in the $k = 1$ case. $A$ is constructed as above, except that in $A_{11}$ and $A_{22}$, we change the off-diagonal entries from 1 to $1.5^2 = 2.25$. Again we can scale everything by a constant so that all entries are in the interval $[1, 2]$, which yields a distance matrix.

Let $B_1 = uv^T$ be the best rank-1 approximation of $B$, where $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^r$. As above we let $\bar{v}$ denote $n/r$ copies of $v$ concatenated to form a vector in $\mathbb{R}^n$. Consider the rank-1 approximation of $A$ given by $A' = [\bar{v}\ u][\bar{v}\ u]^T$. The error on $A_{12}$ and $A_{21}$ is optimal by construction. We need to show that the error on $A_{11}$ and $A_{22}$ is at most $\epsilon \cdot O(n^2)$. To this end we use the fact in our hard distribution that generated $B$ in Section 4, for all supported $B$, the top left and right singular vectors

are nearly the same. Namely, the top-right one is close to $\mathbf{1}$, and the top-left one is close to $1.5 \cdot \mathbf{1}$, for any matrix $B$ in the support.

Concretely, consider an entry in $A_{22}$ whose value is 2.25. The corresponding entry in $A'$ is $v_i v_j$ for $i \neq j$. Each $v_i$ is the mean of a uniformly random vector in $\{1,2\}^r$. Thus it is a scaled binomial random variable with mean 1.5 and variance $1/(4r) = \frac{1}{4}\beta\epsilon$. Furthermore, $v_i$ and $v_j$ are independent. Therefore, the expected squared Frobenius norm error on that entry is

$$\mathbb{E}[(2.25 - v_i v_j)^2] = \mathbf{Var}[v_i v_j] = \mathbf{Var}[v_i] \cdot \mathbf{Var}[v_j] = (\tfrac{1}{4}\beta\epsilon)^2.$$

Thus the expected total error over all of the 2.25 entries (of which there are $O(n^2)$) is $\epsilon \cdot O(n^2)$. The concentration for the 1-entries is even stronger. This completes the proof of the symmetric case for $k = 1$.

# Appendix B. Lower Bound for General $k$

In this section we prove the full statement of Theorem 1.3. The proof largely goes by reduction to the $k = 1$ case, described as follows. We take $k$ copies (referred to as *blocks*) of the hard distribution from the $k = 1$ case, and concatenate them horizontally into an $n \times (kr)$ matrix (where as previously, $r = \Theta(1/\epsilon)$). Then, for each block we pick a Hadamard vector, and add it to all columns in that block. This renders the blocks nearly orthogonal, forcing any low-rank approximation algorithm to compute the majority element of most rows in most blocks, thus solving $\Omega(nk/\epsilon)$ instances of random majority, yielding the desired lower bound. Even though the description is straightforward, the formal proof requires some elaborate technical work, as given in the rest of this section.

Another rather minor difference is that due to having $k$ blocks (which correspond to $k$ clusters of points in the metric space), we cannot add a "heavy row" (which would correspond to a very far point), with all entries set to a large $M > 0$, to each block as we did in the $k = 1$ case. The reason is that the clusters are close (the distance between every two clusters is at most 2), so any point which is far from one cluster must be far from all of them. Thus it would sharpen the spectrum separation of the entire matrix, but not of each block separately, which is the effect we wish to achieve (namely, it would increase the top singular value, but not all top-$k$ singular values.). This is solved by adding $M^2$ light rows instead of a single heavy row. In a light row, we can set all distances to a given cluster $i \in [k]$ to 2, and the rest of the distances to 1. This makes the corresponding point slightly further from cluster $i$ than from the rest of the clusters. Over many similar light rows, this yields the desired effect.

## B.1. Hard distribution

Given $n, k, \epsilon$, let $\beta, C > 0$ be constants that will be chosen later. ($\beta$ will be sufficiently small and $C$ sufficiently large.) Let $r = \beta/\epsilon$, and assume w.l.o.g. this is an integer by letting $\epsilon$ be sufficiently smaller. Let $N = (1 + C)n$.

Next we use the Walsh construction of Hadamard vectors. Recall these are vectors with entries in $\{\pm 1\}^N$ which are pairwise orthogonal. Let $v_1, \ldots, v_k \in \mathbb{R}^N$ be $k$ Hadamard column vectors, which are different than all-1's. We rescale them to have entries in $\{\pm\frac{1}{2}\}$. Let $V^i \in \mathbb{R}^{N \times r}$ be made of $r$ copies of $v^i$ concatenated horizontally.

For every $i = 1, \ldots, k$ let $S^i \in \{\pm\frac{1}{2}\}^{n \times r}$ be made of $n$ vertically stacked instances of majority, after a random permutation of the rows. We complete it to a matrix $\bar{S}^i \in \{0, \pm\frac{1}{2}\}^{N \times r}$ by adding

all-0's lines at the bottom. We use $J$ to denote the all-1's matrix of dimensions implied by context. We form a matrix $\bar{A}^i \in \mathbb{R}^{N \times r}$ by

$$\bar{A}^i = 2J + V^i + \bar{S}^i.$$

We concatenate the $A^i$'s horizontally to obtain a matrix $\bar{A} \in \mathbb{R}^{N \times kr}$. This defines the hard distribution over distance matrices $\bar{A} \in \mathbb{R}^{N \times kr}$.

**Claim B.1** *Every supported $\bar{A}$ is a bipartite distance matrix.*

**Proof** It can checked that all entries of $\bar{A}$ are in $\{1, 1\frac{1}{2}, 2, 2\frac{1}{2}, 3\} \subset [1, 3]$, and thus they satisfy the triangle inequality in a bipartite metric. ■

Let $\bar{B} = \bar{A} - 2J$ and $\bar{B}^i = V^i + \bar{S}^i$. Moreover, let $B \in \mathbb{R}^{n \times kr}$ denote the restriction of $\bar{B}$ to its top $n$ rows. In most of the proof we will actually work with the matrix $\bar{B}$ instead of $\bar{A}$ (cf. Lemma B.5 later on).

### B.1.1. Spectral Properties

**Lemma B.2** *Suppose $(kr)^3 = O(n)$. With probability at least $1 - e^{-\Omega(n/(kr)^{3/2}}$, each of the top $k$ squared singular values of $\bar{B}$ is $\Theta(nr)$, and every other squared singular value is $(\frac{1}{4} \pm o(1))n$.*

**Proof** Similarly to Lemma 4.5, all squared singular value of the random portion of $\bar{B}$ (which are the blocks $S^i$) are $(\frac{1}{4} \pm o(1))n$ with high probability. Since $\bar{B}$ is obtained by adding $k$ rank-1 matrices to that random portion, this bound holds for all but the top-$k$ singular values of $\bar{B}$.

For the upper part of the spectrum, first note that the absolute value of each entry in $\bar{B}$ is distributed uniformly i.i.d. in $\{0, 1\}$. Thus with high probability, $\|\bar{B}\|_F^2 \geq \frac{1}{2}Cnkr$. By the above, the squared singular values except the top-$k$ sum to at most $(\frac{1}{4} + o(1))nkr$ (since there are $kr - k$ of them). Thus the sum of squares of the top-$k$ singular vectors is at least, say, $\frac{1}{4}Cnkr$. On the other hand, for every block $i$, if we multiply $\bar{B}$ on the left by $v^i$ and on the right by the $i$th block indicator, we get $(1 - o(1))Cnr$. Since the Hadamard vectors $\{v_i\}$ are orthogonal and the block indicators are orthogonal, $\bar{B}$ must have at least $k$ singular values whose square is at least $(1 - o(1))Cnr$. The lemma follows. ■

### B.1.2. Bounds on Low-Rank Approximation

For every $i \in [k]$ and $j \in [n]$ let us denote by $s_j^i$ the majority instance which is at the $j$th line of $S^i$. Let $\mu_j^i$ denote its mean. As in the $k = 1$ case, let $\mathbf{1}$ denote the all-1's vector in $\mathbb{R}^r$, and let $P_{\mathbf{1}}$ denote the orthogonal projection on the subspace spanned by it.

**Lemma B.3** $\|\bar{B} - \bar{B}_k\|_F^2 + \epsilon\|\bar{B}\|_F^2 \leq \sum_{i=1}^{k} \sum_{j=1}^{n} \|s_j^i - \mu_j^i \mathbf{1}\|_2^2 + 4(1 + C)\beta nk.$

**Proof** For the first summand, consider the rank-$k$ matrix given by replacing each majority instance $s_j^i$ in $B$ by $\mu_j^i \mathbf{1}$. For the second summand, note that each entry in $\bar{B}$ has magnitude at most 2, thus $\epsilon\|\bar{B}\|_F^2 \leq \epsilon \cdot (1 + C)n \cdot kr = 4(1 + C)\beta nk$. ■

**Corollary B.4** $\|\bar{B} - \bar{B}_k\|_F^2 + \epsilon\|\bar{B}\|_F^2 \leq \frac{1}{4}nkr + 4(1 + C)\beta kn.$

**Proof** In the term $\sum_{i=1}^{n} \|s_i - \mu_i \mathbf{1}\|_2^2$ in the above lemma, if we replace $\mu_i$ by 1.5 (which does not decrease the term since the means are optimal for it), we pay exactly $(\frac{1}{2})^2$ per entry. ■

## B.2. Invoking the Algorithm

Suppose we have a deterministic algorithm that given $\bar{A}$, returns $\bar{A}'$ of rank $k+1$that with probability at least $(2/3) + \delta$ satisfies

$$\|\bar{A} - \bar{A}'\|_F^2 \leq \|\bar{A} - \bar{A}_{k+1}\|_F^2 + \epsilon\|\bar{A}\|_F^2. \tag{13}$$

This the hypothesis of Theorem 1.3, except with $k+1$ instead of $k$; this will be more convenient to work with, and does not change the theorem statement (one can shift $k$ by 1 everywhere).

We now move from working with $\bar{A}$ to working with $\bar{B}$.

**Lemma B.5** *We can obtain an approximation $\bar{B}'$ of rank $k+2$ that satifies*

$$\|\bar{B} - \bar{B}'\|_F^2 \leq \|\bar{B} - \bar{B}_k\|_F^2 + O(\epsilon) \cdot \|\bar{B}\|_F^2. \tag{14}$$

**Proof** We take $\bar{B}' = \bar{A}' - 2J$. Then, for a constant $c > 0$,

$$\begin{aligned}
\|\bar{B} - \bar{B}'\|_F^2 &= \|(\bar{A} - 2J) - (\bar{A}' - 2J)\|_F^2 \\
&= \|\bar{A} - \bar{A}'\|_F^2 \\
&\leq \|\bar{A} - \bar{A}_{k+1}\|_F^2 + \epsilon\|\bar{A}\|_F^2 \\
&\leq \|\bar{A} - \bar{B}_k - 2J\|_F^2 + \epsilon\|\bar{A}\|_F^2 \\
&\leq \|\bar{B} - \bar{B}_k\|_F^2 + c\epsilon\|\bar{B}\|_F^2,
\end{aligned}$$

and we scale $\epsilon$ down by $c$. ■

Combined with Corollary B.4, we get

**Corollary B.6** $\|\bar{B} - \bar{B}'\|_F^2 \leq \frac{1}{4}nkr + 4(1+C)\beta kn$.

Recall that $\bar{B}'_k$ denotes the optimal rank-$k$ approximation of $\bar{B}'$, and thus $\|\bar{B}'_k\|_F^2$ is the sum of squares of the top-$k$ singular values of $\bar{B}'$ (which has a total of $k+2$ singular values).

**Lemma B.7** *The singular values of $\bar{B}'$ satisfy*

$$\sum_{i=1}^{k+2} \left(\sigma_i(\bar{B}) - \sigma_i(\bar{B}')\right)^2 \leq O(C\beta nk).$$

*Furthermore, the two bottom squared singular values are each $O(C\beta nk)$.*

**Proof** Using Lemma B.5 as an upper bound and the Hoffman-Weilandt inequality as a lower bound on $\|\bar{B} - \bar{B}'\|_F^2$,

$$\sum_{i=1}^{k+2} \left(\sigma_i(\bar{B}) - \sigma_i(\bar{B}')\right)^2 + \|\bar{B} - \bar{B}_{k+2}\|_F^2 \leq \|\bar{B} - \bar{B}'\|_F^2 \leq \|\bar{B} - \bar{B}_k\|_F^2 + O(C\beta nk).$$

Observe that $\|\bar{B} - \bar{B}_k\|_F^2 - \|\bar{B} - \bar{B}_{k+2}\|_F^2 = \sigma_{k+1}(\bar{B})^2 + \sigma_{k+2}(\bar{B})^2$, and by Lemma B.2 each of these two summands is $O(n)$, which is less than $O(C\beta nk)$.[7] Plugging this above yields the desired inequality, $\sum_{i=1}^{k+2} \left(\sigma_i(\bar{B}) - \sigma_i(\bar{B}')\right)^2 \leq O(C\beta nk)$.

As for the bottom two singular values of $\bar{B}'$, the inequality just proven yields in particular $\left(\sigma_{k+1}(\bar{B}) - \sigma_{k+1}(\bar{B}')\right)^2 \leq O(C\beta nk)$, hence $\sigma_{k+1}(\bar{B}') \leq \sigma_{k+1}(\bar{B}) + O(\sqrt{C\beta nk})$. As already mentioned above, $\sigma_{k+1}(\bar{B}) = O(\sqrt{n}) = O(\sqrt{C\beta nk})$ by Lemma B.2. Thus $\sigma_{k+1}(\bar{B}')^2 \leq O(C\beta nk)$. The same holds for $\sigma_{k+2}(\bar{B}')$. ■

### B.2.1. Averaging Columns in Blocks

We now carry out the main part of the reduction to the rank-1 case. We do this by showing that $\bar{B}$ can be approximated by the matrix resulting from taking $\bar{B}'$ and replacing each column in each block by the average of columns in that block. Note that in the resulting matrix, each block has rank-1 since its columns are identical. Therefore, by averaging, we could get a rank-1 approximation for a large constant fraction of the blocks. Let us now argue this formally.

Let $\Pi_{\mathbf{1}}$ be the orthogonal projection of $\mathbb{R}^{kr}$ on the subspace spanned by block indicators. Note that for a matrix $Z \in \mathbb{R}^{N \times kr}$, the operation $Z\Pi_{\mathbf{1}}$ averages the columns in each block. The main lemma for this part of this following.

**Lemma B.8** $\|\bar{B} - \bar{B}'\Pi_{\mathbf{1}}\|_F^2 \leq \|\bar{B} - \bar{B}'\|_F^2 + O(\sqrt{C\beta} \cdot nk)$.

The proof will go by showing that the row space of $\bar{B}'$ has to be close to the span of the block indicators, which is the subspace on which $\Pi_{\mathbf{1}}$ projects. (This would yield $\bar{B}' \approx \bar{B}'\Pi_{\mathbf{1}}$ and hence $\|\bar{B} - \bar{B}'\|_F^2 \approx \|\bar{B} - \bar{B}'\Pi_{\mathbf{1}}\|_F^2$, as the lemma asserts). The way we show this is by transitivity, by showing that both subspaces are close to the top-$k$ row space of $\bar{B}$. We will require the following technical linear algebraic claims, whose proofs are deferred to Appendix B.4 for better readability.

**Lemma B.9** *Let $A, B \in \mathbb{R}^{n \times m}$. Let $B = U\Sigma V^T$ be the SVD of $B$. Suppose $\|A - B\|_F^2 \leq \|A\|_F^2 - \Delta$. Then $\|AV\|_F^2 \geq \Delta$.*

**Lemma B.10** *Let $A \in \mathbb{R}^{n \times n}$ and let $A = U\Sigma V^T$ be its SVD. Let $V_k$ be the restriction of $V$ to the top-$k$ right singular vectors of $A$. Let $\Pi$ an orthogonal projection on some $k$-dimensional subspace. If*

$$(1 - \epsilon)k \leq \|V_k^T \Pi\|_F^2 \leq k,$$

*then*

$$\|(A\Pi^{\perp})_k\|_F^2 \leq \|A_{\epsilon k}\|_F^2 + \|(A_{n-k})_k\|_F^2.$$

*(Recall again that $\|X_k\|_F^2$ denotes the sum of squared top-$k$ singular values for every matrix $X$.)*

As a small digression, let us preview that we will use this lemma twice, on the matrices $\bar{B}$ and $\bar{B}'$. In both cases the projection would be $\Pi_{\mathbf{1}}$. In the former case the bound yielded by the lemma would be $O(nk)$, and it the latter case it would be (the better bound) $O(C\beta nk)$. That is, we will get $\|(\bar{B}\Pi^{\perp})_k\|_F^2 = O(nk)$ and $\|(\bar{B}'\Pi^{\perp})_k\|_F^2 = O(C\beta nk)$ (see Appendix B.4 for an elaboration why).

7. We recall that $\beta$ and $C$ are constants that will eventually be chosen such that $C\beta$ is smaller than a sufficiently small constant. It holds that $n = O(C\beta nk)$ is $k$ is larger than a sufficiently large constant, which we can assume w.l.o.g. since we have already proven the $k = 1$ case.

However we still need to establish the condition $(1-\epsilon)k \le \|V_k^T \Pi\|_F^2 \le k$ for both invocations, which we will do shortly.

**Lemma B.11** *Let $V, U, W \in \mathbb{R}^{n\times k}$ matrices such that each has orthonormal columns. Suppose $\|V^T U\|_F^2 \ge (1-\epsilon)k$ and $\|U^T W\|_F^2 \ge (1-\epsilon)k$. Then $\|V^T W\|_F^2 \ge (1-O(\epsilon))k$.*

We now prove Lemma B.8. Let $\bar{B} = U\Sigma V^T$ denote the SVD of $\bar{B}$. Write it as $\bar{B} = U\Sigma V^T = U_T \Sigma_T V_T^T + U_B \Sigma_B V_B^T$ where $\Sigma_T$ are the top $k$ singular values and $\Sigma_B$ are the remaining (bottom) singular values.

**Lemma B.12** *Let $\bar{B}^* \in \mathbb{R}^{N\times kr}$ be a rank-$k'$ matrix such that $\|\bar{B} - \bar{B}^*\|_F^2 \le \frac{1}{4}nkr + O(nk)$. Let $W \in \mathbb{R}^{k'\times kr}$ be an orthonormal basis for the row span of $\bar{B}^*$. Then $\|W V_T\|_F^2 \ge k - O(\epsilon(k+k'))$.*

**Proof** On one hand, since $\|\bar{B}\|_F^2 \ge \frac{1}{2}nkr - O(kn)$, the hypothesis $\|\bar{B} - \bar{B}^*\|_F^2 \le \frac{1}{4}nkr + O(kn)$ implies, by Lemma B.9, $\|\bar{B}W^T\|_F^2 \ge \frac{1}{4}nkr - O(kn)$. On the other hand,

$$\begin{aligned}
\tfrac{1}{4}nkr - O(kn) \le \|\bar{B}W^T\|_F^2 &= \|U_T\Sigma_T V_T^T W^T + U_B \Sigma_B V_B^T W^T\|_F^2 \\
&= \|\Sigma_T V_T^T W^T\|_F^2 + \|\Sigma_B V_B^T W^T\|_F^2 \\
&\le \|\Sigma_T\|_2^2 \|V_T^T W^T\|_F^2 + \|\Sigma_B\|_2^2 \|V_B^T W^T\|_F^2 \\
&\le (\tfrac{1}{4}nr + O(n)) \cdot \|V_T^T W^T\|_F^2 + O(n) \cdot \|V_B^T W^T\|_F^2 \\
&\le (\tfrac{1}{4}nr + O(n)) \cdot \|V_T^T W^T\|_F^2 + O(n) \cdot k'.
\end{aligned}$$

The lemma follows by rearranging and recalling that $r = \Theta(1/\epsilon)$. ∎

We apply the above lemma twice: once with $\bar{B}^*$ being $\bar{B}'$ (whose rank is $k+2$), and once with $\bar{B}^*$ being the matrix obtained from averaging the columns in each block of $\bar{B}$ (note that this matrix has rank is $k$). Since $\Pi_{\mathbf{1}}$ is the orthogonal projection on the row space of that matrix, then by the latter application of Lemma B.12 we have

$$k \ge \|V_T^T \Pi_{\mathbf{1}}\|_F^2 \ge k - O(\epsilon k), \tag{15}$$

which establishes the condition of Lemma B.10, yielding

$$\|(\bar{B}\Pi_{\mathbf{1}}^{\perp})_k\|_F^2 \le O(nk).$$

For the former application, let $\bar{B}' = U'\Sigma'(V')^T$ denote the SVD of $\bar{B}'$. Corollary B.6 provides the requirement of Lemma B.12, which in turn yields $\|(V')^T V_T\|_F^2 \ge k - O(\epsilon k)$. Together with Equation (15), by transitivity (Lemma B.11),

$$k \ge \|(V')^T \Pi_{\mathbf{1}}\|_F^2 \ge k - O(\epsilon k),$$

which establishes the condition of Lemma B.10, yielding

$$\|(\bar{B}'\Pi_{\mathbf{1}}^{\perp})_k\|_F^2 \le O(C\beta nk).$$

Together with the above,

$$\|(\bar{B}\Pi_{\mathbf{1}}^{\perp})_k\|_F \|(\bar{B}'\Pi_{\mathbf{1}}^{\perp})_k\|_F \le O(\sqrt{C\beta}\cdot nk). \tag{16}$$

We can extend this from rank-$k$ to rank-$(k+2)$ since the additional two square singular value of each of the matrices is $O(C\beta nk)$ (cf. Lemmas B.2 and B.7).[8] Since $\bar{B}'\Pi_{\mathbf{1}}^{\perp}$ has rank $k+2$, then by Hoffman-Weilandt,

$$\begin{aligned}
\|\bar{B}\Pi_{\mathbf{1}}^{\perp} - \bar{B}'\Pi_{\mathbf{1}}^{\perp}\|_F^2 &\geq \sum_i \left(\sigma_i(\bar{B}\Pi_{\mathbf{1}}^{\perp}) - \sigma_i(\bar{B}'\Pi_{\mathbf{1}}^{\perp})\right)^2 && \text{by Hoffman-Weilandt} \\
&= \|\bar{B}\Pi_{\mathbf{1}}^{\perp}\|_F^2 - \sum_{i=1}^{k+2} \sigma_i(\bar{B}\Pi_{\mathbf{1}}^{\perp})\sigma_i(\bar{B}'\Pi_{\mathbf{1}}^{\perp}) + \|\bar{B}'\Pi_{\mathbf{1}}^{\perp}\|_F^2 && \operatorname{rank}(\bar{B}'\Pi_{\mathbf{1}}^{\perp}) = k+2 \\
&\geq \|\bar{B}\Pi_{\mathbf{1}}^{\perp}\|_F^2 - \sum_{i=1}^{k+2} \sigma_i(\bar{B}\Pi_{\mathbf{1}}^{\perp})\sigma_i(\bar{B}'\Pi_{\mathbf{1}}^{\perp}) \\
&\geq \|\bar{B}\Pi_{\mathbf{1}}^{\perp}\|_F^2 - \|(\bar{B}\Pi_{\mathbf{1}}^{\perp})_k\|_F \|(\bar{B}'\Pi_{\mathbf{1}}^{\perp})_k\|_F && \text{by Cauchy-Schwartz} \\
&\geq \|\bar{B}\Pi_{\mathbf{1}}^{\perp}\|_F^2 - O(\sqrt{C\beta}\cdot nk) && \text{by Equation (16) .}
\end{aligned}$$

Finally, by Pythagorean identities,

$$\begin{aligned}
\|\bar{B} - \bar{B}'\Pi_{\mathbf{1}}\|_F^2 &= \|\bar{B}\Pi_{\mathbf{1}} - \bar{B}'\Pi_{\mathbf{1}}\|_F^2 + \|\bar{B}\Pi_{\mathbf{1}}^{\perp}\|_F^2 \\
&= \|\bar{B} - \bar{B}'\|_F^2 + \left(\|\bar{B}\Pi_{\mathbf{1}}^{\perp}\|_F^2 - \|\bar{B}\Pi_{\mathbf{1}}^{\perp} - \bar{B}'\Pi_{\mathbf{1}}^{\perp}\|_F^2\right) \\
&\leq \|\bar{B} - \bar{B}'\|_F^2 + O(\sqrt{C\beta}\cdot nk).
\end{aligned}$$

This proves Lemma B.8.

### B.2.2. Relevant Blocks

**Lemma B.13** *There is a subset $I \subset [k]$ of size at least $|I| \geq 0.99k$ such that for every $i \in I$,*

$$\|\bar{B}^i - (\bar{B}')^i P_{\mathbf{1}}\|_F^2 \leq \|\bar{B}^i = (\bar{B}^i)_1\|_F^2 + O(\sqrt{C\beta}\cdot n), \tag{17}$$

*where $(\bar{B}^i)_1$ is (as usual) the optimal rank-1 approximation of $\bar{B}^i$. We refer to blocks $B_i$ with $i \in I$ as* relevant blocks.

**Proof** By Lemma B.8, $\|\bar{B} - \bar{B}'\Pi_{\mathbf{1}}\|_F^2 \leq \|\bar{B} - \bar{B}'\|_F^2 + O(\sqrt{C\beta}\cdot nk)$. Note that the left-hand side equals $\sum_{i=1}^k \|\bar{B}^i - (\bar{B}')^i P_{\mathbf{1}}\|_F^2$. As for the right-hand side, by Equation (14) we have $\|\bar{B} - \bar{B}'\|_F^2 \leq \|\bar{B} - \bar{B}_k\|_F^2 + O(\epsilon)\cdot\|B\|_F^2$, and we recall that $\|B\|_F^2 = O(Cnkr) = O(C\beta nk/\epsilon)$. Furthermore, $\|\bar{B} - \bar{B}_k\|_F^2 \leq \sum_{i=1}^k \|\bar{B}^i - (\bar{B}^i)_1\|_F^2$. Putting it all together yields $\sum_{i=1}^k \|\bar{B}^i - (\bar{B}')^i P_{\mathbf{1}}\|_F^2 \leq \sum_{i=1}^k \|\bar{B}^i - (\bar{B}^i)_1\|_F^2 + O(\sqrt{C\beta}\cdot nk)$, or rearranging,

$$\sum_{i=1}^k \left(\|\bar{B}^i - (\bar{B}')^i P_{\mathbf{1}}\|_F^2 - \|\bar{B}^i - (\bar{B}^i)_1\|_F^2\right) \leq O(\sqrt{C\beta}\cdot nk).$$

Each term in the sum on the left-hand side is non-negative, by the optimality of $(\bar{B}^i)_1$ for rank-1 approximation of $\bar{B}^i$. Therefore we can use an averaging argument (Markov's inequality) and

8. Recall again that we set $C\beta < 1$.

conclude that at least $0.99k$ of the $k$ summands on the left-hand side are at most $100/k$ times the right-hand side. The lemma follows. ■

Fix $i \in I$. Since $(\bar{B}')^i P_{\mathbf{1}}$ is a rank-1 matrix we can write it as $\bar{a}^i (b^i)^T$ where $\bar{a}^i \in \mathbb{R}^N$ and $b^i \in \mathbb{R}^r$. Let $a^i \in \mathbb{R}^n$ denote the restriction of $\bar{a}^i$ to the first $n$ entries. Consider $\|\bar{B}^i - \bar{a}^i (b^i)^T\|_F^2$. Note that the last $Cn$ rows of are either all 1 or all $-1$, depending on the Hadamard vector $v^i$. Let $\sigma_j^i \in \{\pm 1\}$ denote the sign of row $j$. Then the contribution of the last $Cn$ rows is $\sum_{j=n+1}^{Cn} \|\sigma_j^i \mathbf{1} - a_j^i b^i\|_2^2$ which can be rewritten as $\sum_{j=n+1}^{Cn} \|\mathbf{1} - \sigma_j^i a_j^i b^i\|_2^2$. Pick the $j$ that minimizes the term $\|\mathbf{1} - \sigma_j^i a_j^i b^i\|_2^2$ and set all entries $a_{n+1}^i, \ldots, a_{Cn}^i$ to $a_j^i$ with the appropriate sign, to obtain a vector $\hat{a}^i$. By choice of $j$ we have

$$\|\bar{B}^i - \hat{a}^i (b^i)^T\|_F^2 \leq \|\bar{B}^i - \bar{a}^i (b^i)^T\|_F^2. \tag{18}$$

Furthermore,

$$\|\bar{B}^i - \hat{a}^i (b^i)^T\|_F^2 = \|B^i - a^i (b^i)^T\|_F^2 + Cn\|b^i - \mathbf{1}\|_2^2. \tag{19}$$

Combining Equations (18) and (19),

$$\|B^i - a^i (b^i)^T\|_F^2 + Cn\|b^i - \mathbf{1}\|_2^2 \leq \|\bar{B}^i - \bar{a}^i (b^i)^T\|_F^2. \tag{20}$$

If we use Lemma B.13 as an upper bound on $\|\bar{B}^i - \bar{a}^i (b^i)^T\|_F^2$ and Lemma 4.5 as a lower bound on $\|B^i - a^i (b^i)^T\|_F^2$, we get $Cn\|\mathbf{1} - b^i\|_2^2 \leq O(n)$, which rearranges to

$$\|\mathbf{1} - b^i\|_2^2 \leq \frac{O(1)}{C}. \tag{21}$$

This implies Lemmas 4.10 and 4.11 for every relevant block, by the same proofs as their original proofs.

### B.3. Solving Majority

Recall we have a total of $nk$ majority instances (each of length $r$) embedded in $\bar{B}$. Note by the construction of $\bar{B}$, each of them has alphabet either $\{0, 1\}$ or $\{0, -1\}$, depending on the sign of the corresponding entry of the Hadamard vector $v^i$, where $i$ the block in which the instance is embedded.

By Lemma 4.4 and Markov's inequality, at least $0.9nk$ of the instances are typical. For an instance with alphabet $\{0, 1\}$, we solve it using $\bar{B}'$ by reporting that the majority element is 1 if the average over the corresponding entries in $\bar{B}'$ is larger than 0.5, and reporting 0 if it is smaller than 0.5. Instances with alphabet $\{0, -1\}$ are solved similarly with threshold $-0.5$. Note that the solution procedure compares the threshold to the mutual value of the corresponding entries of $\bar{B}'\Pi_{\mathbf{1}}$. If we are correct on an instance, we say it is *solved*, and otherwise *unsolved*. For relevant block $i \in I$, let $S_i'$ denote the subset of majority instances which are both typical and unsolved. Let $S' = \cup_{i \in I} S_i'$ be the subset of all instances which are typical, unsolved, and embedded in a relevant block.

Our goal is to show that we solve each instance with probability at least $2/3$. Since the instances were placed in $\bar{B}$ by random permutation, every instance has the same probability $p$ to be solved, thus we need to show $p \geq 2/3$. Suppose by contradiction that $p < 2/3$. Since at least $0.9nk$ instances are typical, and at least $0.99k$ blocks are relevant, then there is a fixed constant $\zeta > 0$ (this was $\frac{1}{15}$ in the $k = 1$ case) such that $|S'| \geq \zeta nk$.

For every $i \in I$ we have (by definition of relevant blocks),

$$\|\bar{B}^i - (\bar{B}'\Pi_{\mathbf{1}})^i\|_F^2 \leq \|\bar{B}^i - (\bar{B}^i)_1\|_F^2 + O(\sqrt{C\beta}\cdot n) \leq \sum_{j=1}^{n} \|s_j - \mu_j \mathbf{1}\|_2^2 + O(\sqrt{C\beta}\cdot n).$$

By bounding $\sum_{j=1}^{n} \|s_j - \mu_j \mathbf{1}\|_2^2$ in the same way as in the $k = 1$ case,

$$\|\bar{B}^i - (\bar{B}'\Pi_{\mathbf{1}})^i\|_F^2 \leq |S_i'|(\tfrac{1}{4}r - \gamma^2) + \sum_{s_j \notin S_i'} \|s_j - \mu_j \mathbf{1}\|_2^2 + O(\sqrt{C\beta}\cdot n). \tag{22}$$

Similarly, if we denote $(\bar{B}')^i P_{\mathbf{1}} = \bar{a}^i (b^i)^T$ (since this is a rank-1 matrix), then as in the $k = 1$ case,

$$\|B^i - (B'\Pi_{\mathbf{1}})^i\|_F^2 = \sum_{i=1}^{n} \|s_i - \bar{a}_j^i (b^i)^T\|_2^2 \geq |S_i'|(\tfrac{1}{4}r + \eta^2 - 2\gamma\eta) + \sum_{s_j \notin S_i'} \|s_j - \mu_j \mathbf{1}\|_2^2. \tag{23}$$

(We remark that the latter inequality relies on Lemmas 4.10 and 4.11, which were proven in the previous section for relevant blocks, based on eq. (21); as per Lemma 4.10, $\eta = \Theta(1/\sqrt{C})$.)

Together,

$$|S_i'|(\tfrac{1}{4}r + \eta^2 - 2\gamma\eta) \leq |S_i'|(\tfrac{1}{4}r - \gamma^2) + O(\sqrt{C\beta}\cdot n).$$

We sum this over all $i \in I$, and recall that $\sum_{i\in I} |S_i'| = |S'|$. This yields,

$$|S'|(\tfrac{1}{4}r + \eta^2 - 2\gamma\eta) \leq |S'|(\tfrac{1}{4}r - \gamma^2) + O(\sqrt{C\beta}\cdot kn),$$

which rearranges to $|S'|(\gamma - \eta)^2 \leq O(\sqrt{C\beta}\cdot nk)$. Recalling that $|S'| \geq \zeta nk$, we get $\zeta(\gamma - \eta)^2 \leq O(\sqrt{C\beta}\cdot)$. Since $\zeta$ and $\gamma$ are fixed constant, we can take $\eta$ and $\beta$ to be sufficiently small, and arrive at the desired contradiction. ■

## B.4. Deferred Proofs from Appendix B.2.1

**Lemma B.14** *Let $A, B \in \mathbb{R}^{n\times m}$. Let $B = U\Sigma V^T$ be the SVD of $B$. Suppose $\|A - B\|_F^2 \leq \|A\|_F^2 - \Delta$. Then $\|AV\|_F^2 \geq \Delta$.*

**Proof**

$$\begin{aligned}
\|AV\|_F^2 &= \|AVV^T\|_F^2 \\
&= \|A\|_F^2 - \|A(I - VV^T)\|_F^2 && \text{Pythagorean theorem} \\
&\geq \Delta + \|A - B\|_F^2 - \|A(I - VV^T)\|_F^2 \\
&= \Delta + \|AVV^T - BVV^T\|_F^2 + \|A(I - VV^T)\|_F^2 - \|A(I - VV^T)\|_F^2 && \text{Pythagorean theorem} \\
&\geq \Delta.
\end{aligned}$$

■

**Lemma B.15** *Let $A \in \mathbb{R}^{n\times n}$ and let $A = U\Sigma V^T$ be its SVD. Let $V_k$ be the restriction of $V$ to the top-$k$ right singular vectors of $A$. Let $\Pi$ an orthogonal projection on some $k$-dimensional subspace. If*

$$(1-\epsilon)k \le \|V_k^T\Pi\|_F^2 \le k,$$

*then*

$$\|(A\Pi^\perp)_k\|_F^2 \le \|A_{\epsilon k}\|_F^2 + \|(A_{n-k})_k\|_F^2.$$

*(Recall again that $\|X_k\|_F^2$ denotes the sum of squared top-$k$ singular values for every matrix $X$.)*

**Proof** We have $\|(A\Pi^\perp)_k\|_F^2 = \|(A_{n-k}\Pi^\perp + A_k\Pi^\perp)_k\|_F^2$. We can write $A\Pi^\perp$ as $A_{n-k}\Pi^\perp + A_k\Pi^\perp$, and for any vector $x$, $A_{n-k}\Pi^\perp x$ and $A_k\Pi^\perp x$ are orthogonal, and so $\|A\Pi^\perp x\|_2^2 = \|A_{n-k}\Pi^\perp x\|_2^2 + \|A_k\Pi^\perp x\|_2^2$. It follows that

$$\|(A\Pi^\perp)_k\|_F^2 \le \|A_k\Pi^\perp\|_F^2 + \|(A_{n-k}\Pi^\perp)_k\|_F^2.$$

Note that $\|(A_{n-k}\Pi^\perp)_k\|_F^2 \le \|(A_{n-k})_k\|_F^2$. Thus it remains to show $\|A_k\Pi^\perp\|_F^2 \le \|A_{\epsilon k}\|_F^2$, or equivalently, by the Pythagorean theorem, $\|A_k\Pi\|_F^2 \ge \|A_k\|_F^2 - \|A_{\epsilon k}\|_F^2 = \sum_{i=\epsilon k+1}^{k}\sigma_i^2(A)$.

We have $\|A_k\Pi\|_F^2 = \|\Sigma_k V_k^T\Pi\|_F^2$ and we know $\|V_k^T\Pi\|_F^2 \ge k(1-\epsilon)$. Let $R$ be an $n\times n$ rotation matrix that takes $V_k^T$ to $[I_k\ 0]$, where here $I_k$ is the identity matrix of order $k$ and $0$ is an $k\times(n-k)$ zero matrix. Replace $\Pi$ with $R^T\Pi$. Then $\|A_k\Pi\|_F^2 = \|\Sigma_k(V_k^TR)(R^T\Pi)\|_F^2$ and $\|(V_k^TR)(R^T\Pi)\|_F^2 \ge k(1-\epsilon)$. Thus, we can assume w.l.o.g. that $V_k^T = [I_k\ 0]$, and so $R^T\Pi$ has the form $[\Phi\ 0]$, where $\Phi$ is $k\times k$.

Thus $\|A_k\Pi\|_F^2 = \|\Sigma_k\Phi\|_F^2$ subject to $\|\Phi\|_F^2 \ge k(1-\epsilon)$. Also each row of $\Phi$ has squared norm at most 1 since it is a submatrix of a rotation matrix. Consequently, since $\Sigma_k$ is a diagonal matrix, $\|\Sigma_k\Phi\|_F^2$ is minimized when placing all mass of $\Phi$ on the bottom $k(1-\epsilon)$ rows, and in this case it is exactly $\sum_{i=\epsilon k+1}^{k}\sigma_i^2(A)$. ■

We have applied this lemma in Appendix B.2.1 to both $\bar{B}_k$ and $\bar{B}'$. Let us show the resulting upper bound $\|A_{\epsilon k}\|_F^2 + \|(A_{n-k})_k\|_F^2$ in each case.

For $\bar{B}_k$, by Lemma B.2 we know that the top $k$ squared singular values of $\bar{B}$ are $\Theta(Cnr) = O(Cn\beta/\epsilon)$ each, thus $\|\bar{B}_{\epsilon k}\|_F^2 = O(\epsilon k\cdot Cn\beta/\epsilon) = O(C\beta nk)$, and the rest of the squared singular values are $\Theta(n)$, thus $\|(\bar{B}_{n-k})_k\|_F^2 = O(kn)$. The total bound is $O(nk)$.

For $\bar{B}'$,

$$\begin{aligned}
\|\bar{B}'_{\epsilon k}\|_F^2 &= \sum_{i=1}^{\epsilon k}\sigma_i(\bar{B}')^2 \\
&= \sum_{i=1}^{\epsilon k}\left(\sigma_i(\bar{B}') - \sigma_i(\bar{B}) + \sigma_i(\bar{B})\right)^2 \\
&\le \sum_{i=1}^{\epsilon k} 2\left(\sigma_i(\bar{B})^2 + (\sigma_i(\bar{B}) - \sigma_i(\bar{B}'))^2\right) && \text{Similarly to Claim 2.3} \\
&= 2\|\bar{B}_{\epsilon k}\|_F^2 + 2\sum_{i=1}^{\epsilon k}(\sigma_i(\bar{B}) - \sigma_i(\bar{B}'))^2.
\end{aligned}$$

The first term was already upper bounded by $O(C\beta nk)$ above, and the second sum is upper bounded by $O(C\beta nk)$ by Lemma B.7. The term $\|(\bar{B}'_{n-k})_k\|_F^2$ is $O(C\beta nk)$ since there are two remaining eigenvalues and each is $O(C\beta nk)$ by Lemma B.7. The total bound is $O(C\beta nk)$.

**Lemma B.16** *Let $V, U, W \in \mathbb{R}^{n\times k}$ matrices such that each has orthonormal columns. Suppose $\|V^TU\|_F^2 \geq (1-\epsilon)k$ and $\|U^TW\|_F^2 \geq (1-\epsilon)k$. Then $\|V^TW\|_F^2 \geq (1-O(\epsilon))k$.*

**Proof** Note we can replace $V^T$ with $V^TR$ and $U$ with $R^TU$ and $W$ with $R^TW$, where $R$ is an $n \times n$ rotation which takes $U$ to the top $k$ standard unit vectors. All norms in the premise and goal of the claim are preserved. So we can assume that $\|V_{top}\|_F^2 \geq k(1-\epsilon)$, $\|W_{top}\|_F^2 \geq k(1-\epsilon)$, and need to show $\|V^TW\|_F^2 \geq k - O(k\epsilon)$, where "top" means the top $k \times k$ submatrix with remaining rows replaced with 0s. Let $W = W_{top} + W_{rest}$ and $V = V_{top} + V_{rest}$.

Then,

$$\begin{aligned}
\|V^TW\|_F^2 &= \|V_{top}^TW_{top} + V_{rest}^TW_{rest}\|_F^2 \\
&\geq \|V_{top}^TW_{top}\|_F^2 - 2Tr(W_{rest}^TV_{rest}V_{top}^TW_{top}) && \text{(i)} \\
&= \|V_{top}^TW_{top}\|_F^2 - 2Tr(W_{top}W_{rest}^TV_{rest}V_{top}^T) && \text{(ii)} \\
&\geq \|V_{top}^TW_{top}\|_F^2 - 2\|W_{top}W_{rest}\|_F\|V_{rest}V_{top}^T\|_F && \text{(iii)} \\
&\geq \|V_{top}^TW_{top}\|_F^2 - 2\|W_{top}\|_2\|W_{rest}\|_F\|V_{top}\|_2\|V_{rest}\|_F && \text{(iv)} \\
&\geq \|V_{top}^TW_{top}\|_F^2 - 2\|W_{rest}\|_F\|V_{rest}\|_F && \text{(v)} \\
&= \|V_{top}^TW_{top}\|_F^2 - 2(1-\|W_{top}\|_F^2)^{1/2}(1-\|V_{top}\|_F^2)^{1/2} \\
&= \|V_{top}^TW_{top}\|_F^2 - 2(\epsilon k)^{1/2}(\epsilon k)^{1/2} \\
&= \|V_{top}^TW_{top}\|_F^2 - 2\epsilon k, && (*)
\end{aligned}$$

where,

- (i) is by expanding the square and dropping a non-negative term;
- (ii) is by cyclicity of trace;
- (iii) is since $Tr(AB) <= \|A\|_F\|B\|_F$;
- (iv) is by submultiplicativity of operator and Frobenius norm;
- (v) is since $W$ and $V$ are orthonormal so their operator norm is 1, and operator norm does not decrease by taking submatrices.

So we just need to lower bound $\|V_{top}^TW_{top}\|_F^2$, and we can drop the last $n-k$ rows of $V_{top}$ and $W_{top}$ since they are zeros. Next, we write $V_{top}$ in its SVD, $V_{top} = A\Sigma B^T$. Then since $\|V_{top}\|_F^2 \geq k(1-\epsilon)$ and $\|V_{top}\|_2^2 \leq 1$ (since it is a submatrix of $V$), necessarily, there are at least $k(1-2\epsilon)$ singular values of squared value at least $1-2\epsilon$. Indeed, otherwise $\|V_{top}\|_F^2 \leq k(1-2\epsilon)(1-2\epsilon) + 2\epsilon k \cdot 1 < k(1-\epsilon)$ for $\epsilon$ less than a small enough constant. Let $\Sigma_h$ be these singular values and $\Sigma_l$ be the remaining ones. Then

$$\|V_{top}^TW_{top}\|_F^2 = \|\Sigma B^TW_{top}\|_F^2 \geq (1-2\epsilon)\|B_h^TW_{top}\|_F^2.$$

Now $\|W_{top}\|_F^2 \geq k(1-\epsilon)$, and $B_h$ is a $k(1-2\epsilon)$-dimensional subspace of span$(e_1, ..., e_k)$ (the standard unit vectors), and so we can extend it with an orthonormal basis $B'$ so that span$(B_h, B') =$ span$(e_1, ..., e_k)$. Then by the Pythagorean theorem

$$k(1-\epsilon) \leq \|W_{top}\|_F^2 = \|B_h^T W_{top}\|_F^2 + \|B' W_{top}\|_F^2,$$

and since $\|W_{top}\|_2^2 \leq 1$, we have $\|B' W_{top}\|_F^2 \leq \|B'\|_F^2 \leq 2\epsilon k$. Consequently, $\|B_h^T W_{top}\|_F^2 \geq k(1-\epsilon) - 2\epsilon k = k - 3\epsilon k$. Hence, $\|V_{top}^T W_{top}\|_F^2 \geq (1-2\epsilon)k(1-3\epsilon) \geq k(1-O(\epsilon))$. Plugging into (*) gives us our desired $k(1-O(\epsilon))$ lower bound on $\|V^T W\|_F^2$. ■