The Sparse Matrix-Based Random Projection: An Analysis of Matrix Sparsity for Classification

Anonymous authors
Paper under double-blind review

Abstract

In the paper, we study the sparse $\{0, \pm 1\}$-matrix based random projection, which has been widely applied in classification to reduce the data dimension. For such kind of sparse matrices, it is of computational interest to explore the minimum number of nonzero entries ± 1 that supports achieving the best or nearly best classification performance. To achieve this, we analyze the impact of matrix sparsity on the ℓ_1 distance between projected data points. The analysis is inspired by the principle component analysis, which says that the larger distance between projected data points should better capture the variation among original data, and then yield better classification performance. Theoretically, the ℓ_1 distance between projected data points is not only related to the sparsity of sparse matrices, but also to the distribution of original data. Without loss of generality, we consider two typical data distributions, the Gaussian mixture distribution and the two-point distribution, which have been widely used to model the distributions of real data. With the two data distributions, we estimate the ℓ_1 distance between projected data points. It is found that the sparse matrices with only one or at most dozens of nonzero entries per row, can provide comparable or even larger ℓ_1 distances than other more dense matrices, under the matrix size $m \geq \mathcal{O}(\sqrt{n})$. Accordingly, the similar performance trend should also be obtained in classification. This is confirmed with classification experiments on real data of different types, including the image, text, gene and binary quantization data.

1 Introduction

Random projection is an important unsupervised dimensional reduction technique that simply projects high-dimensional data to low-dimensional subspaces by multiplying the data with random matrices (Johnson & Lindenstrauss, 1984). The projection can approximately preserve the pairwise ℓ_2 distance between original data points, or say preserving the structure of original data, thus applicable to classification (Bingham & Mannila, 2001; Fradkin & Madigan, 2003; Wright et al., 2009). To achieve the ℓ_2 distance preservation property, random projection matrices need to follow certain distributions, such as Gaussian matrices (Dasgupta & Gupta, 1999) and sparse $\{0, \pm 1\}$-ternary matrices (shortly called sparse matrices hereafter) (Achlioptas, 2003). In practical applications, sparse matrices are preferred for its much lower complexity both in storage and computation. Considering random projection is often applied to computationally-intensive large-scale classification tasks, it is highly desirable to minimize its complexity. For this purpose, we propose to explore the minimum number of nonzero entries ± 1 that enables the projected data to achieve the best or nearly best classification performance. To the best of our knowledge, no previous study has investigated the problem.

Existing research on random projection is mainly devoted to exploring the distribution of random matrices that well holds the distance preservation property, more precisely, keeping the *expectation* of the pairwise distance between original data points unchanged after random projection and rendering the *variance* relatively small (Dasgupta & Gupta, 1999; Achlioptas, 2003). For the sparse matrix with entries properly scaled, it has been proved that the distance preservation property holds in ℓ_2 norm (Achlioptas, 2003; Li et al., 2006), but *not* in ℓ_1 norm (Brinkman & Charikar, 2003; Li, 2007). Here it is noteworthy that although the ℓ_2 distance preservation property enables random projection to be applied in classification, it can hardly be used to analyze the impact of matrix sparsity (namely the number of nonzero entries) on the follow-on

classification, since the classification accuracy depends on the discrimination between projected data points, rather than the invariance of data structure. For instance, it has been proved that the ℓ_2 distance preservation property tends to become worse as the matrix becomes sparser (Li et al., 2006), namely containing fewer nonzero entries ± 1. However, empirically, it is observed that the sparser matrix structure does not mean a worse classification performance; and usually, very sparse matrices, such as the ones with only one or dozens of nonzero entries per row, can achieve comparable or even better classification performance than other more dense matrices. For this intriguing performance, in the paper we provide reasonable theoretical explanations by analyzing the *variation* of the ℓ_1 distance between projected data points. By the early research of principle component analysis (PCA) (Jolliffe, 2002), it is known that the projection over a *larger* principle component will yield *larger* pairwise distances (equivalently, larger variances) for projected data points, while the larger distance tends to *better* capture the variation (i.e. statistical information) of original data (Jolliffe & Cadima, 2016), and then provide *better* classification performance (Turk & Pentland, 1991).

In the sparse matrix based random projection, the ℓ_1 distance between projected data points is related not only to the sparsity of random matrices, but also to the distribution of original high-dimensional data. To analyze the impact of matrix sparsity on the ℓ_1 distance, we need to first model the distribution of original data. Without loss of generality, we consider two typical data distributions, the Gaussian mixture distribution and the two-point distribution. The former has been widely used to model the distribution of natural data (Torralba & Oliva, 2003; Weiss & Freeman, 2007) or their sparse transforms (Wainwright & Simoncelli, 1999; Lam & Goodman, 2000), while the latter can be used to model the distribution of binary data, such data often occurring in various quantization tasks (Gionis et al., 1999; Hubara et al., 2016; Yang et al., 2019). Benefiting from the two general distributions, as shown later, our theoretical analysis results can be applied to a variety of real data.

Given the two data distributions, by varying the sparsity of sparse matrices, we analyze the *expected* ℓ_1 distance between projected data points and obtain the following two results: 1) The maximum distance tends to be achieved by the sparse matrices with only one nonzero entry per row, as the discrimination between two classes of original data is sufficiently high; 2) The distance tends to converge to a constant value with the increasing of matrix sparsity, and relatively small convergence errors can be achieved when sparse matrices contain only dozens of nonzero entries per row. To summarize, the two results imply that the sparse matrices with only one or at most dozens of nonzero entries per row, perform comparably or even better than the other more dense matrices, in the task of enlarging the expected ℓ_1 distance between projected data points. Accordingly, these matrices should also exhibit similar performance trends on classification. Note that the above analysis is built upon the *expectation* of ℓ_1 distance. To enable the expected distance to be approximated by an actual matrix of size $m \times n$, we need $m \geq \mathcal{O}(\sqrt{n})$. In the experiments, we verify the performance advantage of the sparse matrices mentioned above by conducting classification experiments on real data of different types, including the image, text, gene and binary quantization data. The major contributions of the work can be summarized as follows:

- For the sparse $\{0, \pm 1\}$-matrices based random projection, we for the first time investigate the impact of matrix sparsity on classification, by analyzing random projection from the viewpoint of *distance variation* rather than the conventional *distance preservation*. The proposed analysis is inspired by the early research of PCA (Jolliffe & Cadima, 2016; Turk & Pentland, 1991), that is the larger distance between projected data points should better account for the variation among original data and then yield better classification performance.

- By theoretical and numerical analysis, it is found that the sparse matrices with only one or at most dozens of nonzero entries per row, tend to achieve comparable or even better classification performance than the other more dense matrices, if the original data has the Gaussian mixture distribution or two-point distribution, and the matrices have size $m \geq \mathcal{O}(\sqrt{n})$. This implies that we can drastically reduce the complexity of random projection matrices without losing, or even improving the classification performance.

- The above analysis results are perfectly verified by simulations and experiments. The high *consistency* between theory and practice can be attributed to the good generalizability of the two aforementioned distributions we have adopted to model the original data, which has been recog-

nized in the modeling of various types of data (Torralba & Oliva, 2003; Weiss & Freeman, 2007; Wainwright & Simoncelli, 1999; Lam & Goodman, 2000).

2 Problem Formulation

Consider the random projection of two data points $\boldsymbol{h}, \boldsymbol{h'} \in \mathbb{R}^n$ over a sparse random matrix $\boldsymbol{R} \in \{0, \pm 1\}^{m \times n}$. For the matrix $\boldsymbol{R}$, we attempt to estimate the minimum number of nonzero entries ± 1 that enables maximizing the expected ℓ_1 distance $\mathbb{E}\|\boldsymbol{Rx}\|_1$ between the projections of $\boldsymbol{h}$ and $\boldsymbol{h'}$, where $\boldsymbol{x} = \boldsymbol{h} - \boldsymbol{h'}$. As discussed before, the maximum $\mathbb{E}\|\boldsymbol{Rx}\|_1$ is expected to provide the best classification performance. To determine the minimum sparsity, we need to investigate the changing trend of $\mathbb{E}\|\boldsymbol{Rx}\|_1$ against the varying sparsity of $\boldsymbol{R}$. It can seen that the estimation depends on the distributions of the matrix $\boldsymbol{R}$ and the data $\boldsymbol{h}$. So in the following, we first model the distributions of sparse matrices and real data, and then give the ℓ_1 estimation model.

Notation. Throughout the work, we typically denote a matrix by a bold upper-case letter $\boldsymbol{R} \in \mathbb{R}^{m \times n}$, a vector by a bold lower-case letter $\boldsymbol{r} = (r_1, r_2, ..., r_n)^\top \in \mathbb{R}^n$, and a scalar by a lower-case letter r_i or r. Sometimes, we use the bold letter $\boldsymbol{r}_i \in \mathbb{R}^n$ to denote the i-th row of $\boldsymbol{R} \in \mathbb{R}^{m \times n}$. For ease of presentation, we defer all proofs to Appendix A.

2.1 The distribution of sparse matrices

The sparse random matrix $\boldsymbol{R}$ we aim to study is specified in Definition 1, which has the parameter k counting the number of nonzero entries per row, and is simply called k-sparse to distinguish between the matrices of different sparsity. Instead of the form $\boldsymbol{R} \in \{0, \pm 1\}^{m \times n}$, in the definition we introduce a scaling parameter $\sqrt{\frac{n}{mk}}$ to make the matrix entries have zero mean and unit variance. With this distribution, the matrix will hold the ℓ_2 distance preservation property, that is, keeping the expected ℓ_2 distance between original data points unchanged after random projection (Achlioptas, 2003). Note that the scaling parameter can be omitted in practical applications for easier computation; and the omitting will not change the relative distances between projected data points, thus not affecting the follow-up classification performance.

Definition 1 (k-sparse random matrix)**.** A k-sparse random matrix $\boldsymbol{R} \in \{0, \pm\sqrt{\frac{n}{mk}}\}^{m \times n}$ is defined to be of the following structure properties:

- its each row vector $\boldsymbol{r} \in \{0, \pm\sqrt{\frac{n}{mk}}\}^n$ contains exactly k nonzero entries, $1 \leq k \leq n$;
- the positions of k nonzero entries are arranged uniformly at random;
- each nonzero entry takes the bipolar values $\pm\sqrt{\frac{n}{mk}}$ with equal probability.

2.2 The distribution of original data

For the original high-dimensional data $\boldsymbol{h} \in \mathbb{R}^n$, as discussed before, we investigate two typical distributions, the two-point distribution and the Gaussian mixture distribution. Considering the expected ℓ_1 distance $\mathbb{E}\|\boldsymbol{Rx}\|_1$ is directly related to the pairwise difference $\boldsymbol{x}$ between two original data $\boldsymbol{h}$ and $\boldsymbol{h'}$, namely $\boldsymbol{x} = \boldsymbol{h} - \boldsymbol{h'} = (x_1, x_2, \ldots, x_n)^\top$, we describe the distribution of $\boldsymbol{x}$ for the original data $\boldsymbol{h}$ with the given two distributions.

2.2.1 Two-point distribution

Suppose that the two high-dimensional data $\boldsymbol{h}, \boldsymbol{h'} \in \{\mu_1, \mu_2\}^n$ have their each entry independently following a two-point distribution, where μ_1 and μ_2 are two arbitrary constants. Then the difference $\boldsymbol{x}$ between $\boldsymbol{h}$ and $\boldsymbol{h'}$ has its each entry x_i independently following a ternary discrete distribution

$$x_i \sim \mathcal{T}(\mu, p, q) \tag{1}$$

with the probability mass function $t \in \{-\mu, 0, \mu\}$ under the probabilities $\{q, p, q\}$, where $\mu = \mu_1 - \mu_2$ and $p + 2q = 1$.

2.2.2 Gaussian mixture distribution

When the two data $\boldsymbol{h}$, $\boldsymbol{h}' \in \mathbb{R}^n$ have their each entry independently following a Gaussian mixture distribution, the difference $\boldsymbol{x} = \boldsymbol{h} - \boldsymbol{h}'$ remains a Gaussian mixture (Andrews & Mallows, 1974), which allows each entry x_i to be modeled as

$$x_i \sim \mathcal{M}(\mu, \sigma^2, p, q) \tag{2}$$

with the probability density function

$$f(t) = pf_{\mathcal{N}}(t; 0, \sigma^2) + qf_{\mathcal{N}}(t; \mu, \sigma^2) + qf_{\mathcal{N}}(t; -\mu, \sigma^2) \tag{3}$$

where $f_{\mathcal{N}}(t; \mu, \sigma^2)$ denotes the density function of $t \sim \mathcal{N}(\mu, \sigma^2)$, and the parameters are subject to p, $q \geq 0$ and $p + 2q = 1$.

2.3 The ℓ_1 distance estimation model

With the distributions defined for the original data points $\boldsymbol{h}$, $\boldsymbol{h}' \in \mathbb{R}^n$ and the k-sparse random matrix $\boldsymbol{R} \in \{0, \pm\sqrt{\frac{n}{mk}}\}^{m\times n}$, our goal is to analyze the changing of the expected ℓ_1 distance $\mathbb{E}\|\boldsymbol{R}\boldsymbol{x}\|_1$ (with $\boldsymbol{x} = \boldsymbol{h} - \boldsymbol{h}'$) against varying matrix sparsity k, and determine the minimum sparsity k that corresponds to the maximum or nearly maximum $\mathbb{E}\|\boldsymbol{R}\boldsymbol{x}\|_1$. Notice that we have $\mathbb{E}\|\boldsymbol{R}\boldsymbol{x}\|_1 = m\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}|$, since each row $\boldsymbol{r} \in \mathbb{R}^n$ of $\boldsymbol{R}$ follows an independent and identical distribution by Definition 1. This equivalence suggests that $\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}|$ will share the same changing trend with $\mathbb{E}\|\boldsymbol{R}\boldsymbol{x}\|_1$, when varying k. Then for ease of analysis, instead of $\mathbb{E}\|\boldsymbol{R}\boldsymbol{x}\|_1$, in the following we choose to investigate the changing of $\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}|$ against varying k.

3 The ℓ_1 Distance Estimation with Two-Point Distributed Data

In this section, we investigate the changing of $\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}|$ against varying matrix sparsity k, provided that the original data $\boldsymbol{h}$, $\boldsymbol{h}'$ are drawn from two-point distributions, such that their difference $\boldsymbol{x} = \boldsymbol{h} - \boldsymbol{h}'$ has i.i.d. entries $x_i \sim \mathcal{T}(\mu, p, q)$, as specified in (1).

3.1 Theoretical analysis

Theorem 1. Let $\boldsymbol{r}$ be a row vector of a k-sparse random matrix $\boldsymbol{R} \in \{0, \pm\sqrt{\frac{n}{mk}}\}^{m\times n}$, and $\boldsymbol{x} \in \mathbb{R}^n$ with i.i.d. entries $x_i \sim \mathcal{T}(\mu, p, q)$. It can be derived that

$$\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}| = 2\mu\sqrt{\frac{n}{mk}}\sum_{i=0}^{k} C_k^i p^i q^{k-i} \left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil} \tag{4}$$

and

$$\mathrm{Var}(|\boldsymbol{r}^\top\boldsymbol{x}|) = \frac{2q\mu^2 n}{m} - \frac{4\mu^2 n}{mk}\left(\sum_{i=0}^{k} C_k^i p^i q^{k-i} \left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil}\right)^2 \tag{5}$$

where C_k^i is a binomial coefficient $\binom{k}{i}$ and $\lceil\alpha\rceil = \min\{\beta : \beta \geq \alpha, \beta \in \mathbb{Z}\}$. By (4), $\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}|$ satisfies the following two properties:

(P1) When $p \leq 0.188$, $\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}|$ has its maximum at $k = 1$.

(P2) When $k \to \infty$, $\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}|$ converges to a constant:

$$\lim_{k\to\infty} \frac{\sqrt{m}}{\mu\sqrt{n}}\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}| = 2\sqrt{q/\pi}, \tag{6}$$

which has the convergence error for finite k upper-bounded by

$$\left|\frac{\sqrt{m}}{\mu\sqrt{n}}\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}| - 2\sqrt{q/\pi}\right| \leq \frac{\sqrt{\pi} + \sqrt{2}}{\sqrt{\pi k}}. \tag{7}$$

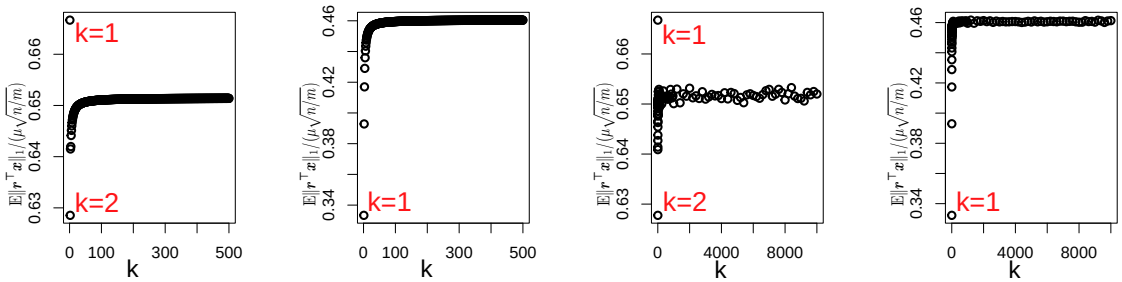


Figure 1: The value of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ calculated by (4) with $p = 1/3$ (a) and $p = 2/3$ (b), and estimated by statistical simulation with $p = 1/3$ (c) and $p = 2/3$ (d), provided $x_i \sim \mathcal{T}(\mu, p, q)$, $\mu = 1$.

Remarks of Theorem 1: In P1 and P2, we characterize the changing trends of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ against varying matrix sparsity k, which are further discussed as follows.

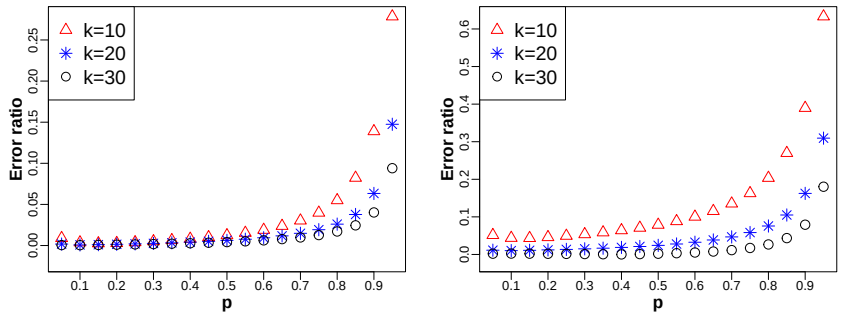
- By P1, $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ can achieve its maximum value at $k = 1$, if the probability p of $x_i = 0$ is sufficiently small (≤ 0.188). This condition means that the difference $\boldsymbol{x}$ between two data points $\boldsymbol{h}$ and $\boldsymbol{h}'$ should have a sufficient number of nonzero entries, and also implies that the two data $\boldsymbol{h}$ and $\boldsymbol{h}'$ should be sufficiently distinguishable from each other. Then we can say that for two discriminative data distributions, the best classification performance should be able to be achieved using very sparse random matrices with sparsity $k = 1$, by virtue of the maximum $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ achieved at $k = 1$.
- By P2, $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ will converge to a constant that depends on the data distribution and matrix size, as k tends to infinity. Note that in (6) we describe the convergence with $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ instead of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$, in terms of the fact that both formulas share the same changing trend against varying k, but the former has fewer parameters, only involving k and p. Moreover, it is noteworthy that the convergence error, namely the difference between the values of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ with finite k and infinite k, is upper-bounded in (7), and the bound indicates a convergence speed $\mathcal{O}(1/\sqrt{k})$. By the bound (7), it is easy to further derive that

$$\frac{\left|\frac{\sqrt{m}}{\mu\sqrt{n}}\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| - 2\sqrt{q/\pi}\right|}{2\sqrt{q/\pi}} \leq \eta, \quad \text{if} \quad k \geq \frac{(\sqrt{\pi}+\sqrt{2})^2}{4q\eta^2} \tag{8}$$

  where η can be an arbitrary positive constant. It is seen that η establishes an upper bound for the ratio between the convergence error with the convergence value (called the convergence ratio error for short); and for any given upper bound η, there exists a lower bound for sparsity k to hold it. This means that within the bound of k derived for small η, $\mathbb{E}|r^\top x|$ will take similar values for different k, and accordingly, the different k should yield similar classification performance. Then we can say that the sparse matrices with small k (taking the values around its lower bound), will provide comparable classification performance with the other more dense matrices with larger k. Note that the lower bound of k derived in (8) contains slack, and in practice it tends to be much smaller, as demonstrated in the following numerical analysis.

3.2 Numerical analysis

To more closely examine the changing trends of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ against varying matrix sparsity k (derived in P1 and P2), we directly compute the value of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ by (4). Note that besides the parameter k, $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ also involves the parameter p, the probability of $x_i = 0$ as specified in (1). So we investigate $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ over $k \in [1, 500]$ for different $p \in (0, 1)$. For brevity, we here only provide the results of $p = 1/3$ and $2/3$ in Figs. 1 (a) and (b), see the supplement for more results. By the numerical analysis results, we revisit the two changing trends described in P1 and P2 and obtain more positive conclusions:



(a) (b)

Figure 2: The convergence error ratios of three different $k \in \{10, 20, 30\}$ over varying p are derived for two-point distributed data (a) and Gaussian mixture data (b), by computing the left side of the inequality of η shown respectively in Eqs. (8) and (14). The parameters involved in computation are set as introduced in the corresponding numerical analysis.

(P3) When $p \leq 1/3$, such as the case of $p = 1/3$ shown in Fig. 1(a), $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ tends to achieve its maximum value at $k = 1$, but at other larger k when $p > 1/3$, such as the case of $p = 2/3$ illustrated in Fig. 1(b). Compared to the theoretical result P1, the numerical result relaxes the upper bound of p from ≤ 0.188 to $\leq 1/3$, enlarging the data space that allows the maximum $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ to be reached at $k = 1$. More precisely, the above relaxed condition requires each entry x_i of $\boldsymbol{x}$ to be nonzero with a probability greater than 2/3, instead of a probability greater than 0.812 (as required by P1). This superficially reduces the demand for the amount of nonzero entries occuring in the difference vector $\boldsymbol{x}$ between two data points $\boldsymbol{h}$ and $\boldsymbol{h}'$, and essentially, reduces the requirement for the discrimination between $\boldsymbol{h}$ and $\boldsymbol{h}'$.

(P4) With the increasing of k, as the two cases of $p = 1/3$ and 2/3 shown in Figs. 1(a) and (b), $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ tends to converge to the limit value $2\sqrt{q/\pi}$ derived in (6), where $q = (1-p)/2$. Furthermore, it can be seen that the convergence speed is fast, allowing small convergence errors to be reached with small k, typically in the range of a few tens. For instance, in Fig. 2(a) we derive the convergence error ratios as defined in (8), which give the values close to zero when $k \geq 20$ and p is relatively small. Recall that the small p implies the case that the original data have high discrimination. With the decreasing of data discrimination, we should need larger k to achieve small convergence errors.

Besides the expectation $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ of pairwise distances as discussed above, the variance $\mathrm{Var}(|\boldsymbol{r}^\top \boldsymbol{x}|)$ of pairwise distances derived in (5) is also a factor that may affect the classification performance. By computing (5), interestingly, we find that with the increasing of k, $\mathrm{Var}(|\boldsymbol{r}^\top \boldsymbol{x}|)$ exhibits a changing trend opposite to that of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$, see the supplement for details. In other words, the larger expectation corresponds to the smaller variance. Considering both larger expectations and smaller variances are favorable to classification, we can say that the two factors achieve consistent results in estimating the classification performance.

3.3 Statistical simulation

To verify the correctness of Theorem 1, including the expression (4) of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ and its two properties P1 and P2, we here estimate the expectation value $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ (against varying k) by performing averaging over the statistically generated samples of $\boldsymbol{r}$ and $\boldsymbol{x}$. If the theorem results are correct, the statistical simulation results should be consistent with the numerical analysis results P3 and P4 (derived by Theorem 1). The simulation is introduced as follows. First, we randomly generate 10^6 pairs of $\boldsymbol{r}$ and $\boldsymbol{x}$ from their respective distributions, i.e. $\boldsymbol{r} \in \{0, \pm\sqrt{\frac{n}{mk}}\}^n$ with k nonzero entries randomly distributed, and $\boldsymbol{x}$ with i.i.d. $x_i \sim \mathcal{T}(\mu, p, q)$. Then, the average value of $|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ is derived as the final estimate of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$. The parameters for the distributions of $\boldsymbol{r}$ and $\boldsymbol{x}$ are set as follows: $m = 1$, $n = 10^4$, $\mu = 1$, and $p = 1/3$ or 2/3. The data dimension $n = 10^4$ allows us to increase k from 1 to 10^4. The

average value of $|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ at each k is provided in Figs. 1(c) and (d), respectively for the cases of $p = 1/3$ and $2/3$. Note that the choices of m, n and μ will not affect the changing trend of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ against k. Comparing the numerical analysis results and the simulation results provided in Fig. 1, namely contrasting (a) vs. (c) and (b) vs. (d), it is seen that two kinds of results exhibit similar changing trends for $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$. The similarity between them validates Theorem 1, as well as the numerical analysis results P3 and P4.

Moreover, it is noteworthy that what we estimate is an *expected* distance $\mathbb{E}\|\boldsymbol{R}\boldsymbol{x}\|_1$ (equivalently, $m\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$), rather than the actual distance $\|\boldsymbol{R}\boldsymbol{x}\|_1$ we will derive with a given matrix sample. To approximate the expected distance, by Property 1 the actual matrices should have the size of $m \geq \mathcal{O}(\sqrt{n})$.

Property 1. Let $\boldsymbol{r}_i$ be the i-th row of a k-sparse random matrix $\boldsymbol{R} \in \{0, \pm\sqrt{\frac{n}{mk}}\}^{m\times n}$, and $\boldsymbol{x} \in \mathbb{R}^n$ with i.i.d. entries $x_i \sim \mathcal{T}(\mu, p, q)$. Suppose $z = \frac{1}{m}\|\boldsymbol{R}\boldsymbol{x}\|_1 = \frac{1}{m}\sum_{i=1}^m |\boldsymbol{r}_i^\top \boldsymbol{x}|$. For arbitrary small $\varepsilon, \delta > 0$, we have the probability $\Pr\{|z - \mathbb{E}z| \leq \varepsilon\} \geq 1 - \delta$, if $\frac{m^2}{m+1} \geq \frac{q\mu^2 n}{\varepsilon^2\delta}$; and the condition can be relaxed to $m^2 \geq \frac{2q\mu^2 n}{\varepsilon^2\delta}$, for a given $\boldsymbol{x}$.

4 The ℓ_1 Distance Estimation with Gaussian Mixture Data

In this section, we consider the case that the original data $\boldsymbol{h}$, $\boldsymbol{h}'$ are drawn from Gaussian mixture distributions, such that their difference $\boldsymbol{x} = \boldsymbol{h} - \boldsymbol{h}'$ has i.i.d. entries $x_i \sim \mathcal{M}(\mu, \sigma^2, p, q)$, as specified in (2). With such data, the changing of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ against varying matrix sparsity k is analyzed.

4.1 Theoretical analysis

Theorem 2. Let $\boldsymbol{r}$ be a row vector of a k-sparse random matrix $\boldsymbol{R} \in \{0, \pm\sqrt{\frac{n}{mk}}\}^{m\times n}$, and $\boldsymbol{x} \in \mathbb{R}^n$ with i.i.d. entries $x_i \sim \mathcal{M}(\mu, \sigma^2, p, q)$. It can be derived that

$$\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| = 2\mu\sqrt{\frac{n}{mk}}T_1 + \sigma\sqrt{\frac{2n}{\pi m}}T_2 - 2\mu\sqrt{\frac{n}{mk}}T_3 \tag{9}$$

$$T_1 = \sum_{i=0}^{k} C_k^i p^i q^{k-i} \left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil}$$

$$T_2 = \sum_{i=0}^{k} C_k^i p^i q^{k-i} \sum_{j=0}^{k-i} C_{k-i}^j e^{-\frac{(k-i-2j)^2\mu^2}{2k\sigma^2}}$$

$$T_3 = \sum_{i=0}^{k} C_k^i p^i q^{k-i} \sum_{j=0}^{k-i} C_{k-i}^j \Phi\left(-\frac{|k-i-2j|\mu}{\sqrt{k}\sigma}\right)$$

and

$$\mathrm{Var}(|\boldsymbol{r}^\top \boldsymbol{x}|) = \frac{n}{m}(\sigma^2 + 2q\mu^2) - \left(\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|_1\right)^2 \tag{10}$$

where $\Phi(\cdot)$ is the distribution function of $\mathcal{N}(0, 1)$. Further, we have

$$\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| \leq \mu\sqrt{\frac{n}{m}} + \sigma\sqrt{\frac{2n}{\pi m}}, \tag{11}$$

and

$$\lim_{k\to\infty} \frac{\sqrt{m}}{\mu\sqrt{n}}\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| = \sqrt{\frac{2}{\pi}(\sigma^2 + 2q\mu^2)} \tag{12}$$

which has the convergence error for finite k upper-bounded by

$$\left|\frac{\sqrt{m}}{\mu\sqrt{n}}\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| - \sqrt{2(\sigma^2 + 2q\mu^2)/\pi}\right| \leq \frac{4\sigma^3\left[p + 2q(1 + \mu^2/\sigma^2)^{3/2}\right]}{(\sigma^2 + 2q\mu^2)\sqrt{\pi k}} + \frac{\sqrt{2}[3\sigma^4 + 2q(6\sigma^2\mu^2 + \mu^4)]}{\sqrt{(\sigma^2 + 2q\mu^2)\pi k}}. \tag{13}$$

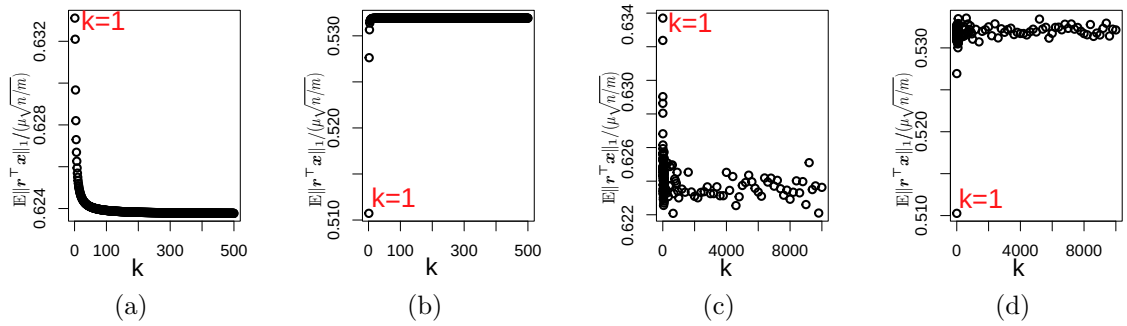


Figure 3: The value of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/\sqrt{n/m}$ calculated by (9) with $p = 1/2$ (a) and $p = 2/3$ (b), and estimated by statistical simulation with $p = 1/2$ (c) and $p = 2/3$ (d), provided $x_i \sim \mathcal{M}(p, q, \mu, \sigma^2)$, $\mu = 1$ and $\sigma = 1/3$.

Remarks of Theorem 2: In Eqs. (12) and (13), we obtain two results similarly as in P2 of Theorem 1. First, $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ converges to a constant with speed $\mathcal{O}(1/\sqrt{k})$. Second, by (13), we can derive the lower bound of k that ensures the convergence error ratio upper-bounded by any given constant η:

$$\frac{\left|\frac{\sqrt{m}}{\mu\sqrt{n}}\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| - \sqrt{2(\sigma^2 + 2q\mu^2)/\pi}\right|}{\sqrt{2(\sigma^2 + 2q\mu^2)/\pi}} \leq \eta, \quad \text{if} \quad k \geq \left(\frac{4\sigma^3[p + 2q(1 + \mu^2/\sigma^2)^{3/2}]}{(\sigma^2 + 2qu^2)^{3/2}\sqrt{2}\eta} + \frac{3\sigma^4 + 2q(6\sigma^2\mu^2 + \mu^4)}{(\sigma^2 + 2q\mu^2)\eta}\right)^2. \tag{14}$$

As discussed in the remarks of Theorem 1, the lower bound of k derived for small η indicates a matrix sparsity k which can provide comparable classification performance with the other larger sparsity k. Usually, as shown in the following numerical analysis, the lower bound of k is small, allowing us to obtain sparse matrices. Moreover, the numerical analysis demonstrates that similarly to P1 of Theorem 1, $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ in (12) also has its maximum attained at $k = 1$, when the data distribution parameter p specified in (2) takes relatively small values.

4.2 Numerical analysis

In this part, we directly compute the value of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ by (9). Note that $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ involves four parameters: k, p, μ, and σ. In computing (9), we fix $\mu = 1$ and vary other parameters in the ranges of $\sigma/\mu \in (0, 1/3)$, $p \in (0, 1)$ and $k \in [1, 500]$. For easy simulation, we here upper bound σ/μ by $1/3$, in view of the fact that σ/μ is usually not large for real data, while larger bounds empirically also work. Empirically, the changing trend of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ is not sensitive to σ/μ, but sensitive to p, namely the probability of each entry x_i of the data difference $\boldsymbol{x}$ taking zero value, as specified in (2). In Figs. 3(a) and (b), we provide two typical results of $p = 1/2$ and $2/3$, and observe two properties similar to the previous P3 and P4:

(P5) When $p \leq 1/2$, such as the case of $p = 1/2$ and $\sigma/\mu = 1/3$ shown in Fig. 3(a), $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/\mu\sqrt{n/m}$ tends to obtain its maximum at $k = 1$, but at other larger k when $p > 1/2$, such as the case of $p = 2/3$ and $\sigma/\mu = 1/3$ shown in Fig. 3(b). It can be seen that the upper bound of p obtained here for Gaussian mixture data is relaxed from $2/3$ to $1/2$ compared to the bound derived in P3 for two-point distributed data. This implies a wider range of data distributions that enables obtaining the maximum $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/\mu\sqrt{n/m}$ at $k = 1$.

(P6) With the increasing of k, as the two results shown in Fig. 3(a) and (b), $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ converges to the limit value derived in (12). Similarly to the convergence discussed in P4 for two-distributed data, the convergence error ratio defined in (14) can approach zero with small k, such as $k = 20$ shown in Fig. 2(b), especially when p is relatively small, namely the original data having relatively high discrimination.

For P5 and P6, their similarity to P3 and P4 is not surprising, since the two-point distribution $x_i \sim \mathcal{T}(\mu, p, q)$ can be viewed as an extreme case of the Gaussian mixture distribution $x_i \sim \mathcal{M}(\mu, \sigma^2, p, q)$ with $\sigma \to 0$.

Thanks to the good generalizability of Gaussian mixture models, as will be seen in our experiments, the two properties analyzed above hold for a variety of real data.

Again note that we should have the matrix row size $m \geq \mathcal{O}(\sqrt{n})$, such that the actual distance $\|\boldsymbol{R}^\top \boldsymbol{x}\|_1$ computed with a single random matrix sample can approximate the expected distance $\mathbb{E}\|\boldsymbol{R}^\top \boldsymbol{x}\|_1$ (equivalently $m\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$) derived with (9). The analysis is similar to Property 1, thus omitted here.

4.3 Statistical simulation

Similarly as in Section 3.3, we here verify the correctness of Theorem 2, including the expression (9) of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ and its convergence (12) by performing statistical simulation on $\boldsymbol{x}$ and $\boldsymbol{r}$. The simulation results should agree with the numerical analysis results P5 and P6, if the theorem is correct. In the simulation, we estimate the value of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/\sqrt{n/m}$ by drawing 10^6 pairs of $\boldsymbol{x}$ and $\boldsymbol{r}$ from their respective distributions and then computing the average of $\|\boldsymbol{r}^\top \boldsymbol{x}\|_1/\sqrt{n/m}$ as the estimate. The parameters of the distributions of $\boldsymbol{x}$ and $\boldsymbol{r}$ are set as follows: $m = 1$, $n = 10000$, $\mu = 1$, $\sigma = 1/3$ and $p = 1/2$ or $2/3$. The data dimension $n = 10000$ allows us to vary k from 1 to 10000. The average value of $|\boldsymbol{r}^\top \boldsymbol{x}|/\sqrt{n/m}$ at each k is presented in Figs. 3(c) and (d), which have $p = 1/2$ and $2/3$, respectively. Comparing the numerical analysis results and the simulation results shown in Fig. 3, namely contrasting (a) vs. (c) and (b) vs. (d), it can be seen that two kinds of results are roughly consistent with each other. The consistency validates Theorem 2, as well as the numerical analysis results P5 and P6.

5 Experiments

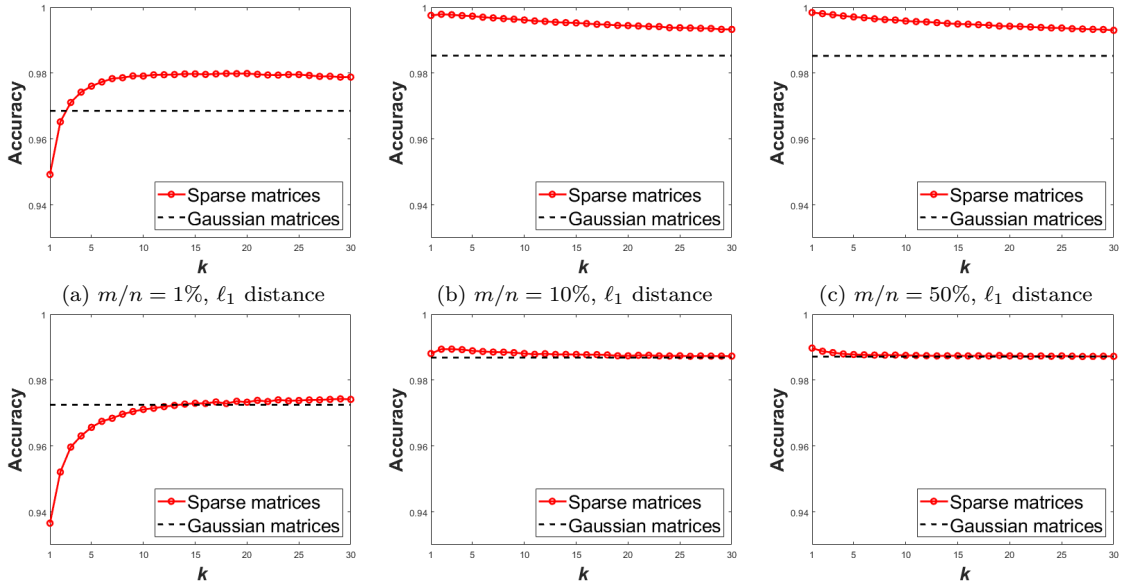
In this section, we aim to verify that the impact of the varying matrix sparsity k on classification is consistent with its impact on the ℓ_1 distance between projected data as analyzed in Theorems 1 and 2; and more precisely, our goal is to demonstrate that the sparse matrices with only one or at most dozens of nonzero entries per row can provide comparable or even better classification performance than other more dense matrices, under the constraint of matrix size $m \geq \mathcal{O}(\sqrt{n})$.

5.1 Data

Without loss of generality, we evaluate four different types of data, including the image dataset YaleB (Georghiades et al., 2001; Lee et al., 2005), the text dataset Newsgroups (Joachims, 1997), the gene dataset AMLALL (Golub et al., 1999) and binary image dataset MNIST (Deng, 2012). The former three kinds of data can be modeled by Gaussian mixtures, while the last one belongs to the two-point distribution. The data settings are introduced as follows. YaleB contains 40×30-sized face images of 38 persons, with about 64 faces per person. Newsgroups consists of 20 categories of 3000-dimensional text data, with 500 samples per category. AMLALL contains 25 samples taken from patients suffering from acute myeloid leukemia (AML) and 47 samples from patients suffering from acute lymphoblastic leukemia (ALL), with each sample expressed with a 7129-dimension gene vector. MNIST involves 10 classes of 28×28-sized handwritten digit images in MNIST, with 6000 samples per class and with each image pixel 0-1 binarized. Note here we reduce the dimension of the data in YaleB and Newsgroups for easy simulation, and this will not influence our comparative study.

5.2 Implementation

The random projection based classification is implemented by first multiplying original data with k-sparse random matrices and then classifying the resulting projections with a classifier. To faithfully reflect the impact of the varying data distance on classification, we adopt the simple nearest neighbor classifier (NNC) (Cover & Hart, 1967) for classification, which has performance absolutely dependent on the pairwise distance between data points, without involving extra operations to improve data discrimination. In fact, our classification performance analysis on matrix sparsity could also be verified with other more sophisticated classifiers, like SVMs (Cortes & Vapnik, 1995), see Appendix B.



(a) $m/n = 1\%$, ℓ_1 distance (b) $m/n = 10\%$, ℓ_1 distance (c) $m/n = 50\%$, ℓ_1 distance

(d) $m/n = 1\%$, ℓ_2 distance (e) $m/n = 10\%$, ℓ_2 distance (f) $m/n = 50\%$, ℓ_2 distance

Figure 4: Classification accuracy of the sparse matrix-based and Gaussian matrices-based random projections for image data (YaleB, DCT features), with varying matrix sparsity $k \in [1, 30]$, three different projection ratios $m/n = 1\%$, 10% and 50%, and two distance metrics ℓ_1 and ℓ_2.

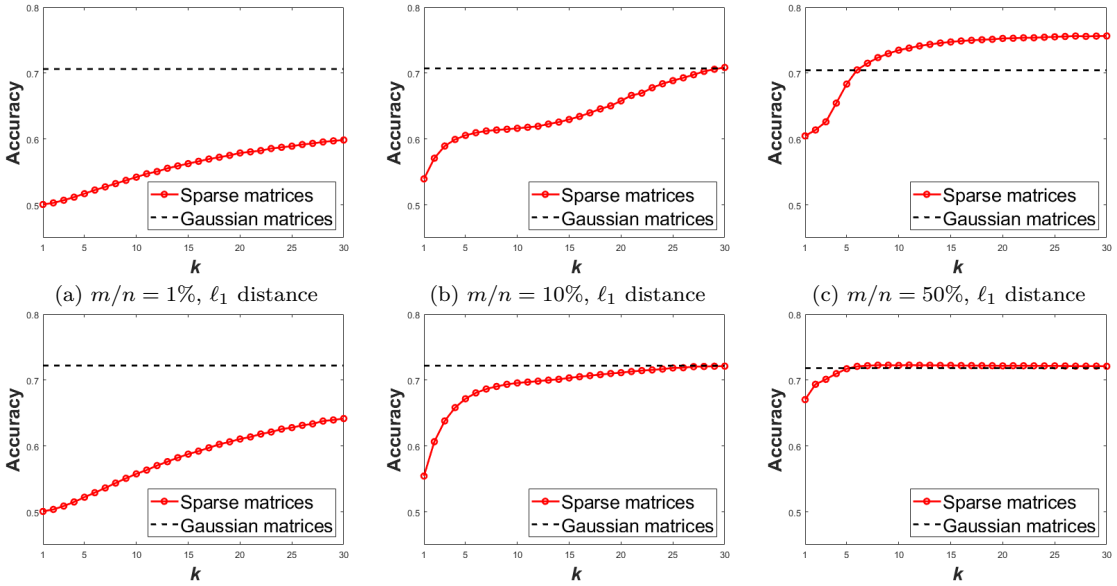
For each dataset, we will enumerate all possible class pairs in it to perform binary classification. In each class, we have one half of samples randomly selected for training and the rest for testing. To suppress the instability of random matrices and obtain relatively stable classification performance, as in (Bingham & Mannila, 2001), we repeat the random projection-based classification 5 times for each sample and make the final classification decision by vote. For comparison, the performance of the Gaussian matrix based random projection is provided. Although the classification performance of sparse matrices is analyzed with ℓ_1 distance, we also test and verify the performance on the popular ℓ_2 distance.

5.3 Results

The classification results of four kinds of data are provided in Figs. 4–7, respectively. For each kind of data, as can be seen, we evaluate the classification performance of sparse matrices with varying sparsity $k \in [1, 30]$, three different projection ratios $m/n = 1\%$, 10% and 50%, and two distance metrics ℓ_1 and ℓ_2. Note that the data dimensions n we test here are on the order of thousands. With such scale of n, it is easy to deduce that the condition of $m \geq \mathcal{O}(\sqrt{n})$ will be satisfied as $m/n = 10\%$ and 50%, but be violated as $m/n = 1\%$.

Let us first examine the case of satisfying $m \geq \mathcal{O}(\sqrt{n})$, namely the cases of $m/n = 10\%$ and 50% as shown in Figs. 4–7(b)(c). It is seen that the four kinds of data all achieve their best performance with relatively small matrix sparsity k (< 30), such as with $k = 1$ in Fig. 4(c) and $k = 15$ in Fig. 5(c). But in the case of $m/n = 1\%$ which violates the condition of $m \geq \mathcal{O}(\sqrt{n})$, as shown in Figs. 4–7(a), the four kinds of data with an exception of AMLALL all fail to reach their top performance within $k < 30$. For AMLALL with $m/n = 1\%$, as illustrated in Fig. 6(a), it fails to get the desired decreasing performance trend and performs poorly at $k = 1$, in contrast to the cases of $m/n = 10\%$ and $m/n = 50\%$ shown in Figs. 6(b)(c). Overall, the experimental results on four different kinds of data all agree with our theoretical analysis: the sparse matrices with only one or at most about dozens of nonzero entries per row, achieve comparable or even better classification performance than other more dense matrices, under the size of $m \geq \mathcal{O}(\sqrt{n})$.

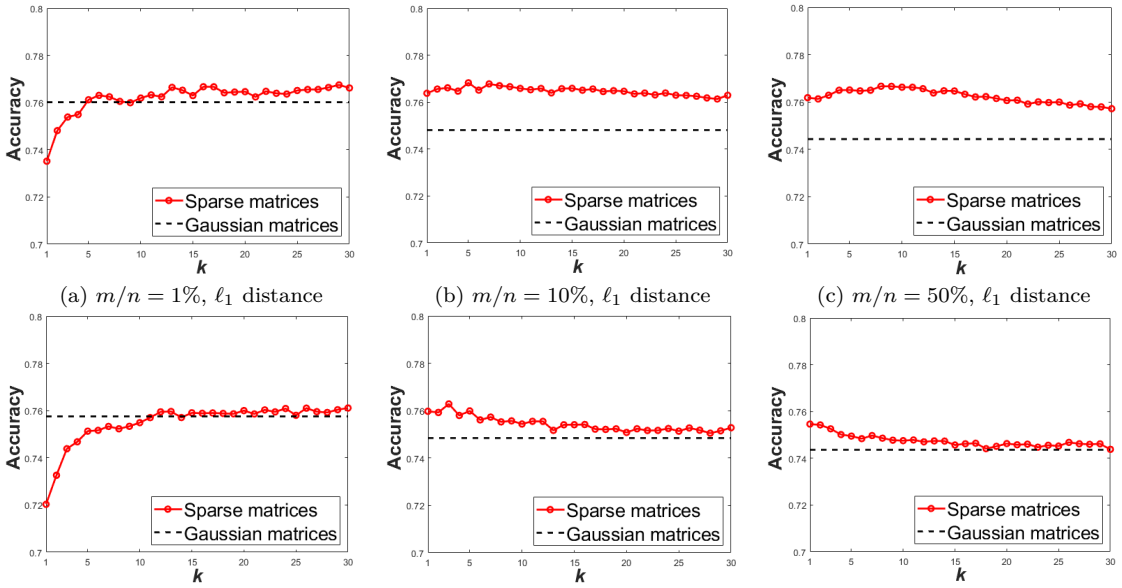
The changing trend of the classification performance against varying matrix sparsity k also consists with our theoretical analysis. More precisely, it can be seen from Figs. 4–7(b)(c) that the classifications of four datasets quickly converge to stable performance with the increasing matrix sparsity k. The difference between them mainly lies in the initial stage of the convergence. Specifically, as illustrated in Figs. 4(b)(c)



(a) $m/n = 1\%$, ℓ_1 distance (b) $m/n = 10\%$, ℓ_1 distance (c) $m/n = 50\%$, ℓ_1 distance

(d) $m/n = 1\%$, ℓ_2 distance (e) $m/n = 10\%$, ℓ_2 distance (f) $m/n = 50\%$, ℓ_2 distance

Figure 5: Classification accuracy of the sparse matrix-based and Gaussian matrix-based random projections for text data (Newsgroups), with varying matrix sparsity $k \in [1, 30]$, three different projection ratios $m/n = 1\%$, 10% and 50%, and two distance metrics ℓ_1 and ℓ_2.

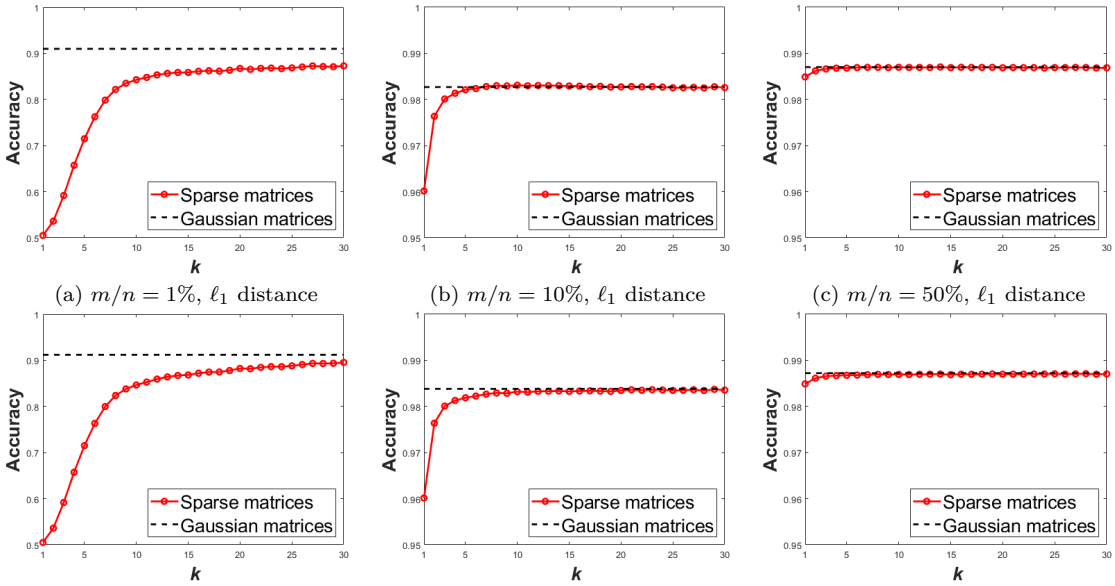


(a) $m/n = 1\%$, ℓ_1 distance (b) $m/n = 10\%$, ℓ_1 distance (c) $m/n = 50\%$, ℓ_1 distance

(d) $m/n = 1\%$, ℓ_2 distance (e) $m/n = 10\%$, ℓ_2 distance (f) $m/n = 50\%$, ℓ_2 distance

Figure 6: Classification accuracy of sparse matrix-based and Gaussian matrix-based random projections for gene data (AMLALL), with varying matrix sparsity $k \in [1, 30]$, three different projection ratios $m/n = 1\%$, 10% and 50%, and two distance metrics ℓ_1 and ℓ_2.

and 6(b)(c), the convergence curves on the datasets YaleB and AMLALL both exhibit the declining trend at the initial increasing region of k, consistent with the numerical analysis result depicted in Fig. 3(a) (discussed in P5 and P6). As for the curves on the other two datasets Newsgroups and MNIST, as shown in Figs. 5(b)(c) and Figs. 7(b)(c), they both exhibit the trend of initially increasing with k, aligning with the numerical analysis results illustrated in Fig. 3(b) (P5 and P6) and Fig. 1(b) (P4 and P5).



(a) $m/n = 1\%$, ℓ_1 distance (b) $m/n = 10\%$, ℓ_1 distance (c) $m/n = 50\%$, ℓ_1 distance

(d) $m/n = 1\%$, ℓ_2 distance (e) $m/n = 10\%$, ℓ_2 distance (f) $m/n = 50\%$, ℓ_2 distance

Figure 7: Classification accuracy of the sparse matrix-based and Gaussian matrix-based random projections for binary image data (MNIST, binarized pixels), with varying matrix sparsity $k \in [1, 30]$, three different projection ratios $m/n = 1\%$, 10% and 50%, and two distance metrics ℓ_1 and ℓ_2.

Although the classification performance of sparse matrices is analyzed with ℓ_1 distance, it can be seen that the performance also holds for ℓ_2 distance, when comparing the upper row and the bottom row results shown in Figs. 4–7. This generalization can be attributed to the closeness of the two metrics (Gionis et al., 1999; Figiel et al., 1977). Moreover, experiments show that sparse matrices perform comparably or even better than the popularly used Gaussian matrices. This allows us to replace Gaussian matrices with sparse matrices, for much lower complexity.

6 Conclusion

For the sparse $\{0, \pm 1\}$-matrix based random projection, we have analyzed the impact of matrix sparsity on classification. It is found that the sparse matrices with only one or at most dozens of nonzero entries per row, can provide comparable or even better classification performance than other more dense matrices, when the matrices have size $m \geq \mathcal{O}(\sqrt{n})$ and the original data are sufficiently discriminative. Moreover, it is empirically observed that the sparse matrices also compare favorably with the popularly used Gaussian matrices, and furthermore, the performance advantage we estimate with ℓ_1 distance also holds with ℓ_2 distance. Theses results imply that our sparse matrices have wide applications. Finally, it is noteworthy that our theoretical analysis exhibits high consistency with the experiments on real data of different types, owing to the good generalizability of the typical data distributions adopted in our statistical analysis.

Besides the contribution to random projection, our classification performance analysis on sparse matrices is helpful to understand the competitive performance of deep ternary networks, which are generated by ternarizing the parameters and/or activations of full-precision networks and enjoy very sparse structures (Li et al., 2016; Zhu et al., 2017; Wan et al., 2018; Marban et al., 2020; Rokh et al., 2023). Despite suffering from significant quantization errors, interestingly, deep ternary networks usually have acceptable performance loss and sometimes can even provide performance gains. The reason for this intriguing phenomenon remains unclear. Considering deep networks can be modeled as a cascade of random projections (Giryes et al., 2016), our analysis of sparse matrix-based random projection can be viewed as a layerwise analysis of deep ternary networks. The sparse ternary matrices we have estimated with good classification performance partly explains the good performance of sparse ternary networks.

References

D. Achlioptas. Database-friendly random projections: Johnson–Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.

David F Andrews and Colin L Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(1):99–102, 1974.

Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 245–250, 2001.

Bo Brinkman and Moses Charikar. On the impossibility of dimension reduction in ℓ_1. *Journal of the ACM*, pp. 766–788, 2003.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

S. Dasgupta and A. Gupta. An elementary proof of the Johnson–Lindenstrauss lemma. *Technical Report, UC Berkeley*, (99–006), 1999.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

T. Figiel, J. Lindenstrauss, and V. D. Milman. The dimension of almost spherical sections of convex bodies. *Acta Mathematica*, 139:53–94, 1977.

Dmitriy Fradkin and David Madigan. Experiments with random projections for machine learning. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 517–522, 2003.

Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE transactions on pattern analysis and machine intelligence*, 23(6):643–660, 2001.

Aristides Gionis, Piotr Indyk, and Rajeev Motwani. Similarity search in high dimensions via hashing. In *Proceedings of the 25th International Conference on Very Large Data Bases*, 1999.

Raja Giryes, Guillermo Sapiro, and Alex M Bronstein. Deep neural networks with random gaussian weights: A universal classification strategy? *IEEE Transactions on Signal Processing*, 64(13):3444–3457, 2016.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016.

Thorsten Joachims. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 143–151, 1997.

W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemp. Math.*, 26:189–206, 1984.

Ian T Jolliffe. *Principal component analysis*. Springer, 2002.

Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065): 20150202, 2016.

Edmund Y Lam and Joseph W Goodman. A mathematical analysis of the DCT coefficient distributions for images. *IEEE transactions on image processing*, 9(10):1661–1666, 2000.

Kuang-Chih Lee, Jeffrey Ho, and David J Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on pattern analysis and machine intelligence*, 27(5):684–698, 2005.

Fengfu Li, Bin Liu, Xiaoxing Wang, Bo Zhang, and Junchi Yan. Ternary weight networks. *arXiv preprint arXiv:1605.04711*, 2016.

P. Li, T. J. Hastie, and K. W. Church. Very sparse random projections. *in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006.

Ping Li. Very sparse stable random projections for dimension reduction in ℓ_α $(0 < \alpha \leq 2)$ norm. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007.

Arturo Marban, Daniel Becking, Simon Wiedemann, and Wojciech Samek. Learning sparse & ternary neural networks with entropy-constrained trained ternarization (EC2T). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 722–723, 2020.

Babak Rokh, Ali Azarpeyvand, and Alireza Khanteymoori. A comprehensive survey on model quantization for deep neural networks in image classification. *ACM Transactions on Intelligent Systems and Technology*, 14(6):1–50, 2023.

Nathan Ross. Fundamentals of stein's method. *Probability Surveys*, 8:210–293, 2011.

Pante Stănică. Good lower and upper bounds on binomial coefficients. *JIPAM. Journal of Inequalities in Pure & Applied Mathematics [electronic only]*, 2, 2001.

Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: computation in neural systems*, 14(3):391–412, 2003.

M.A. Turk and A.P. Pentland. Face recognition using eigenfaces. In *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 586–591, 1991.

Aad W Van der Vaart. *Asymptotic statistics*. Cambridge university press, 2000.

Martin J. Wainwright and Eero P. Simoncelli. Scale mixtures of gaussians and the statistics of natural images. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, 1999.

Diwen Wan, Fumin Shen, Li Liu, Fan Zhu, Jie Qin, Ling Shao, and Heng Tao Shen. TBN: Convolutional neural network with ternary inputs and binary weights. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 315–332, 2018.

Yair Weiss and William T Freeman. What makes a good model of natural images? In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. IEEE, 2007.

J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:210–227, 2009.

J. Yang, X. Shen, J. Xing, X. Tian, H. Li, B. Deng, J. Huang, and X. Hua. Quantization networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Chenzhuo Zhu, Song Han, Huizi Mao, and William J. Dally. Trained ternary quantization. In *International Conference on Learning Representations*, 2017.

A Appendix

A.1 Proof of Theorem 1

Proof. In the following, we sequentially prove (4), (5), P1 and P2.

Proofs of (4) **and** (5)**:** With the distributions of $\boldsymbol{r}$ and $\boldsymbol{x}$, we can write $\|\boldsymbol{r}^\top \boldsymbol{x}\|_1 = \sqrt{\frac{n}{mk}}\mu \left|\sum_{i=1}^{k} z_i\right|$, where $z_i \in \{-1, 0, 1\}$ with probabilities $\{q, p, q\}$. Then, it can be derived that

$$\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| = \mu\sqrt{\frac{n}{mk}} \sum_{i=0}^{k} C_k^i p^i q^{k-i} \sum_{j=0}^{k-i} C_{k-i}^j |k-i-2j|, \tag{15}$$

among which $\sum_{j=0}^{k-i} C_{k-i}^j |k-i-2j|$ can be expressed as

$$\sum_{j=0}^{k-i} (C_{k-i}^j |k-i-2j|) = 2\left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil}, \tag{16}$$

where $\lceil \alpha \rceil = \min\{\beta : \beta \geq \alpha, \beta \in \mathbb{Z}\}$. Combine (15) and (16), we can obtain (4).

Next, we can derive the variance of $|\boldsymbol{r}^\top \boldsymbol{x}|$

$$\begin{aligned} \mathrm{Var}(|\boldsymbol{r}^\top \boldsymbol{x}|) &= \mathrm{Var}(\boldsymbol{r}^\top \boldsymbol{x}) - \left(\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|\right)^2 \\ &= \frac{2q\mu^2 n}{m} - \frac{4\mu^2 n}{mk}\left(\sum_{i=0}^{k} C_k^i p^i q^{k-i} \left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil}\right)^2. \end{aligned} \tag{17}$$

Proof of P1: This part aims to prove

$$\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|_{k=1} > \mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|_{k>1},$$

where the subscript $k = 1$ denotes the case of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ with $k = 1$, and the subscript $k > 1$ means the case of k taking any integer value greater than 1. In the following, we will calculate and compare $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ in terms of the two cases. For the case of $k = 1$, by (4), it is easy to derive that

$$\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|_{k=1} = 2q\mu\sqrt{\frac{n}{m}}. \tag{18}$$

Then, let us see the case of computing $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|_{k>1}$. By (4), $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|_{k>1}$ is the sum of $\frac{2}{\sqrt{k}} C_k^i p^i q^{k-i} \left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil}$ multiplied by $\mu\sqrt{\frac{n}{m}}$. To compute $\frac{2}{\sqrt{k}} C_k^i p^i q^{k-i} \left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil}$, we consider separately two cases: $k - i$ is even or odd, as detailed below.

Case 1: Suppose $k - i$ is even. We have

$$\begin{aligned} &\frac{2}{\sqrt{k}} C_k^i p^i q^{k-i} \left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil} \\ \leq\ &\frac{1}{\sqrt{k}} C_k^i p^i q^{k-i} (k-i) 2^{k-i} \sqrt{\frac{2}{(k-i)\pi}} \\ \leq\ &\sqrt{\frac{2}{\pi}} C_k^i p^i (2q)^{k-i}, \end{aligned} \tag{19}$$

since $C_{2\gamma}^{\gamma} \leq \frac{2^{2\gamma}}{\sqrt{\gamma\pi}}$, where γ is a positive integer (Stănică, 2001).

Case 2: Suppose $k-i$ is odd. We have

$$\begin{aligned}
&\frac{2}{\sqrt{k}} C_k^i p^i q^{k-i} \left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil} \\
\leq &\frac{1}{\sqrt{k}} C_k^i p^i q^{k-i} (k-i) 2^{k-i} \sqrt{\frac{2}{(k-i-1)\pi}} \\
= &\sqrt{\frac{2}{\pi}} C_k^i p^i (2q)^{k-i} \frac{k-i}{\sqrt{k(k-i-1)}}
\end{aligned} \tag{20}$$

Given $k \geq 5$, we further have

$$\frac{k-i}{\sqrt{k(k-i-1)}} < 1 \quad \text{for} \quad 2 \leq i \leq k-2,$$

and for $i = k-1$ or k,

$$\frac{2}{\sqrt{k}} C_k^i p^i q^{k-i} \left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil} < \sqrt{\frac{2}{\pi}} C_k^i p^i (2q)^{k-i}.$$

To sum up, when $k-i$ is odd,

$$\begin{aligned}
&\frac{2}{\sqrt{k}} C_k^i p^i q^{k-i} \left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil} \\
\leq &\begin{cases} \sqrt{\frac{2}{\pi}} C_k^i p^i (2q)^{k-i}, & k \geq 5, i \geq 2, \\ \frac{2}{\sqrt{k}} C_k^i p^i q^{k-i} (k-i) C_{k-i-1}^{\frac{k-i-1}{2}}, & \text{otherwise.} \end{cases}
\end{aligned} \tag{21}$$

According to the results (19) and (21) derived in the above two cases, we know that $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|_{k>1}$ can be computed in terms of two cases, $2 \leq k \leq 4$ and $k \geq 5$. For the case of $2 \leq k \leq 4$, by (4), we have

$$\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| = \begin{cases} \frac{\mu\sqrt{n}}{\sqrt{2m}}(4q^2 + 4pq), & k = 2, \\ \frac{\mu\sqrt{n}}{\sqrt{3m}}(12q^3 + 12pq^2 + 6p^2 q), & k = 3, \\ \frac{\mu\sqrt{n}}{\sqrt{m}}(12q^4 + 24pq^3 + 12p^2q^2 + 4p^3 q), & k = 4, \end{cases} \tag{22}$$

and for the case of $k \geq 5$, with (19) and (21), we have

$$\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| \leq \mu\sqrt{\frac{2n}{\pi m}} + \mu\sqrt{\frac{n}{m}}(2q)^5 \left(\frac{3\sqrt{5}}{8} - \sqrt{\frac{2}{\pi}} \right). \tag{23}$$

By (18), (22) and (23), we can derive that

$$\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|_{k=1} > \mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|_{k>1}$$

holds under the condition of $p \leq 0.188$. Then P1 is proved.

In what follows, we elaborate the proof of (23) by considering two cases of k, being even or odd.

Case 1: Suppose $k \geq 5$ and k is even. Combining (19) and (21), we have

$$\begin{aligned}\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| \leq & \mu\sqrt{\frac{n}{m}} C_k^1 p(2q)^{k-1}\left(\frac{\sqrt{k}}{2^{k-1}} C_{k-1}^{\frac{k}{2}-1} - \sqrt{\frac{2}{\pi}}\right) \\ & + \mu\sqrt{\frac{2n}{\pi m}} \sum_{i=0}^{k} C_k^i p^i (2q)^{k-i}. \end{aligned} \tag{24}$$

Denote $h_1(k) = \frac{\sqrt{k}}{2^{k-1}} C_{k-1}^{\frac{k}{2}-1}$. For

$$\frac{h_1(k+2)}{h_1(k)} = \frac{k+1}{\sqrt{k(k+2)}} > 1$$

we have

$$h_1(k) = \frac{\sqrt{k}}{2^{k-1}} C_{k-1}^{\frac{k}{2}-1} \leq \lim_{k\to\infty} h_1(k) = \sqrt{\frac{2}{\pi}}. \tag{25}$$

Then, it follows from (24) and (25) that

$$\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| \leq \mu\sqrt{\frac{2n}{\pi m}}. \tag{26}$$

Case 2: Suppose $k \geq 5$ and k is odd. Combining (19) and (21), we have

$$\begin{aligned}\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| \leq & \mu\sqrt{\frac{n}{m}} C_k^0 (2q)^{k}\left(\frac{\sqrt{k}}{2^{k-1}} C_{k-1}^{\frac{k-1}{2}} - \sqrt{\frac{2}{\pi}}\right) \\ & + \mu\sqrt{\frac{2n}{\pi m}} \sum_{i=0}^{k} C_k^i p^i (2q)^{k-i}. \end{aligned} \tag{27}$$

Denote $h_2(k) = \frac{\sqrt{k}}{2^{k-1}} C_{k-1}^{\frac{k-1}{2}}$. For

$$\frac{h_2(k+2)}{h_2(k)} = \frac{\sqrt{k(k+2)}}{k+1} < 1$$

we have

$$h_2(k) = \frac{\sqrt{k}}{2^{k-1}} C_{k-1}^{\frac{k-1}{2}} \leq h_2(5) = \frac{\sqrt{5}}{2^4} C_4^2. \tag{28}$$

Then, it follows from (27) and (28) that

$$\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| \leq \mu\sqrt{\frac{2n}{\pi m}} + \mu\sqrt{\frac{n}{m}}(2q)^5\left(\frac{3\sqrt{5}}{8} - \sqrt{\frac{2}{\pi}}\right).$$

Proof of P2: For ease of analysis, we first define the function

$$g(\boldsymbol{r}^\top \boldsymbol{x}; k, p) = \frac{\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|_k}{\mu\sqrt{n/m}} = \mathbb{E}\left|\frac{1}{\sqrt{k}}\sum_{i=1}^{k} z_i\right|, \tag{29}$$

where $\{z_i\}$ is independently and identically distributed and $z_i \in \{-1, 0, 1\}$ with probabilities $\{q, p, q\}$. By the Lindeberg-Lévy Central Limit Theorem, we have

$$\frac{1}{\sqrt{k}}\sum_{i=1}^{k} z_i \rightsquigarrow Z, \tag{30}$$

where $Z \sim N(0, 2q)$.

Then based on (23), we have for $k \geq 5$,

$$\mathbb{E}\left|\frac{1}{\sqrt{k}}\sum_{i=1}^{k} z_i\right| \leq \sqrt{\frac{2}{\pi}} + (2q)^5\left(\frac{3\sqrt{5}}{8} - \sqrt{\frac{2}{\pi}}\right).$$

It means that

$$\lim_{M\to+\infty}\limsup_{k\to+\infty}\mathbb{E}\left[\left|\frac{1}{\sqrt{k}}\sum_{i=1}^{k} z_i\right| 1\left\{\left|\frac{1}{\sqrt{k}}\sum_{i=1}^{k} z_i\right| > M\right\}\right] = 0.$$

Hence, $\left|\frac{1}{\sqrt{k}}\sum_{i=1}^{k} z_i\right|$ is an asymptotically uniformly integrable sequence.

According to Theorem 2.20 in (Van der Vaart, 2000), we obtain

$$\begin{aligned}\lim_{k\to+\infty}\frac{\sqrt{m}}{\mu\sqrt{n}}\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}| &= \lim_{k\to+\infty}\mathbb{E}\left|\frac{1}{\sqrt{k}}\sum_{i=1}^{k} z_i\right| \\ &= \mathbb{E}\,|Z| \\ &= 2\sqrt{\frac{q}{\pi}}.\end{aligned}$$

Next, let us investigate the error of the above convergence with respect to k. Following the definitions and properties described in Eqs. (29) and (30), we further suppose $t_i = \frac{1}{\sqrt{2q}}z_i$ and $Q \sim N(0, 1)$, and get

$$\begin{aligned}&\left|\frac{\sqrt{m}}{\mu\sqrt{n}}\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}| - 2\sqrt{q/\pi}\right| \\ =&\left|\mathbb{E}\left|\frac{1}{k}\sum_{i=1}^{k} z_i\right| - \mathbb{E}|Z|\right| \\ =&\sqrt{2q}\left|\mathbb{E}\left|\frac{1}{k}\sum_{i=1}^{k} t_i\right| - \mathbb{E}|Q|\right| \\ \leq&\sqrt{2q}d_w\left(\mathbb{E}\left|\frac{1}{k}\sum_{i=1}^{k} t_i\right|, \mathbb{E}|Q|\right)\end{aligned}$$

where $d_w(\nu, \upsilon)$ denotes the Kolmogorov metric, with the form

$$\begin{aligned}d_w(\nu,\upsilon) &= \sup_{h\in\mathcal{H}}\left|\int h(x)d\nu(x) - \int h(x)d\upsilon(x)\right|, \\ \mathcal{H} &= \{h : \mathbb{R}\to\mathbb{R} : |h(x) - h(y)| \leq |x - y|\}.\end{aligned}$$

By the Theorem 3.2 in Ross (2011), since $\{t_i\}$ are i.i.d and $\mathbb{E}t_i = 0$, $\mathbb{E}t_i^2 = 1$, $\mathbb{E}|t_i|^4 < \infty$, we have

$$\begin{aligned}d_w\left(\mathbb{E}\left|\frac{1}{k}\sum_{i=1}^{k} t_i\right|, \mathbb{E}|Q|\right) &\leq \frac{1}{k^{3/2}}\sum_{i=1}^{k}\mathbb{E}|t_i|^3 + \frac{\sqrt{2}}{\sqrt{\pi}k}\sqrt{\sum_{i=1}^{k}\mathbb{E}t_i^4} \\ &= \frac{1}{\sqrt{2qk}} + \frac{\sqrt{2}}{\sqrt{2q\pi k}},\end{aligned}$$

and then

$$\left|\frac{\sqrt{m}}{\mu\sqrt{n}}\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}| - 2\sqrt{q/\pi}\right| \leq \frac{\sqrt{\pi}+\sqrt{2}}{\sqrt{\pi k}}.$$

□

A.2 Proof of Property 1

Proof. This problem can be addressed using the Chebyshev's Inequality, which requires us to first derive $\mathbb{E}z$ and $\mathrm{Var}(z)$. Note that $\mathbb{E}z = \mathbb{E}(\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{r}_i^\top \boldsymbol{x}|) = \mathbb{E}(|\boldsymbol{r}_i^\top \boldsymbol{x}|)$ has been derived in (4). In the sequel, we need to first solve $\mathrm{Var}(z) = \mathbb{E}z^2 - (\mathbb{E}z)^2$, which has

$$\begin{aligned}\mathbb{E}z^2 &= \mathbb{E}(\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{r}_i^\top \boldsymbol{x}|)^2 \\ &= \frac{1}{m^2}\mathbb{E}(\sum_{i=1}^{m}|\boldsymbol{r}_i^\top \boldsymbol{x}|^2) + \frac{1}{m^2}\mathbb{E}(\sum_{i\neq j}|\boldsymbol{r}_i^\top \boldsymbol{x}|\cdot|\boldsymbol{r}_j^\top \boldsymbol{x}|) \\ &= \frac{2q\mu^2 n}{m^2} + \frac{m-1}{2m}\mathbb{E}(|\boldsymbol{r}_i^\top \boldsymbol{x}|\cdot|\boldsymbol{r}_j^\top \boldsymbol{x}|).\end{aligned} \tag{31}$$

For the second term in the above result, it holds

$$\mathbb{E}(|\boldsymbol{r}_i^\top \boldsymbol{x}|\cdot|\boldsymbol{r}_j^\top \boldsymbol{x}|) \leq \mathrm{Var}(|\boldsymbol{r}_i^\top \boldsymbol{x}|) + (\mathbb{E}|\boldsymbol{r}_i^\top \boldsymbol{x}|)^2 = \mathrm{Var}(|\boldsymbol{r}_i^\top \boldsymbol{x}|) + (\mathbb{E}z)^2, \tag{32}$$

by the covariance property

$$\begin{aligned}\mathrm{Cov}(|\boldsymbol{r}_i^\top \boldsymbol{x}|, |\boldsymbol{r}_j^\top \boldsymbol{x}|) &= \mathbb{E}(|\boldsymbol{r}_i^\top \boldsymbol{x}|\cdot|\boldsymbol{r}_j^\top \boldsymbol{x}|) - \mathbb{E}|\boldsymbol{r}_i^\top \boldsymbol{x}|\cdot\mathbb{E}|\boldsymbol{r}_j^\top \boldsymbol{x}| \\ &= \rho\sqrt{\mathrm{Var}(|\boldsymbol{r}_i^\top \boldsymbol{x}|)}\cdot\sqrt{\mathrm{Var}(|\boldsymbol{r}_j^\top \boldsymbol{x}|)} \\ &= \rho\mathrm{Var}(|\boldsymbol{r}_i^\top \boldsymbol{x}|),\end{aligned} \tag{33}$$

where $\rho \in (-1, 1)$ is the correlation coefficient.

Substituting (31) into $\mathrm{Var}(z) = \mathbb{E}z^2 - (\mathbb{E}z)^2$, by the inequality (32) and (17), we can derive

$$\begin{aligned}\mathrm{Var}(z) &\leq \frac{2q\mu^2 n}{m^2} + \frac{m-1}{2m}[\mathrm{Var}(|\boldsymbol{r}_i^\top \boldsymbol{x}|) + (\mathbb{E}z)^2] - (\mathbb{E}z)^2 \\ &= \frac{2q\mu^2 n}{m^2} + \frac{m-1}{2m}\cdot\frac{2q\mu^2 n}{m^2} - (\mathbb{E}z)^2 \\ &= \frac{(m+1)q\mu^2 n}{m^2} - (\mathbb{E}z)^2.\end{aligned} \tag{34}$$

With the above inequality about $\mathrm{Var}(z)$, we can further explore the condition that holds the desired probability

$$\Pr\{|z - \mathbb{E}z| \leq \varepsilon\} \geq 1 - \delta. \tag{35}$$

By the Chebyshev's Inequality, (35) will be achieved, if $\mathrm{Var}(z)/\varepsilon^2 \leq \delta$; and according to (34), this condition can be satisfied when $\frac{m^2}{m+1} \geq \frac{q\mu^2 n}{\varepsilon^2\delta}$.

In the above analysis, we consider a random $\boldsymbol{x}$. For a given $\boldsymbol{x}$, the condition of holding (35) can be further relaxed to $m^2 \geq \frac{2q\mu^2 n}{\varepsilon^2\delta}$, since in this case $|\boldsymbol{r}_i^\top \boldsymbol{x}|$ is independent between different $i \in [m]$, such that $\mathrm{Var}(z)$ changes to be (17) divided by m. □

A.3 Proof of Theorem 2

Proof. First, we derive the absolute moment of $z \sim \mathcal{N}(\mu, \sigma^2)$ as

$$\mathbb{E}|z| = \sqrt{\frac{2}{\pi}}\sigma e^{-\frac{\mu^2}{2\sigma^2}} + \mu\left(1 - 2\Phi\left(-\frac{\mu}{\sigma}\right)\right) \tag{36}$$

which will be used in the sequel. With the distributions of $\boldsymbol{r}$ and $\boldsymbol{x}$, we have $|\boldsymbol{r}^\top \boldsymbol{x}| = \sqrt{\frac{n}{mk}} \left| \sum_{i=1}^{k} x_i \right|$. For easier expression, assume $y = \sum_{i=1}^{k} x_i$, then the distribution of y can be expressed as

$$f(y) = \sum_{i=0}^{k} \sum_{j=0}^{k-i} C_k^i C_{k-i}^j p^i q^{k-i} \frac{1}{\sqrt{2\pi k}\sigma} e^{-\frac{(y-(2j+i-s)\mu)^2}{2k\sigma^2}}.$$

Then, by (36) we can derive that

$$\begin{aligned}
\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| = & \sqrt{\frac{n}{mk}} \sum_{i=0}^{k} \sum_{j=0}^{k-i} \Big[C_k^i C_{k-i}^j p^i q^{k-i} \\
& \times \int_{-\infty}^{+\infty} \frac{|y|}{\sqrt{2\pi k}\sigma} e^{-\frac{(y-(2j+i-s)\mu)^2}{2k\sigma^2}} dy \Big] \\
= & \, 2\mu \sqrt{\frac{n}{mk}} \sum_{i=0}^{k} C_k^i p^i q^{k-i} \left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil} \\
& - 2\mu \sqrt{\frac{n}{mk}} \sum_{i=0}^{k} C_k^i p^i q^{k-i} \sum_{j=0}^{k-i} C_{k-i}^j \Phi \left(-\frac{|k-i-2j|\mu}{\sqrt{k}\sigma} \right) \\
& + \sigma \sqrt{\frac{2n}{\pi m}} \sum_{i=0}^{k} C_k^i p^i q^{k-i} \sum_{j=0}^{k-i} C_{k-i}^j e^{-\frac{(k-i-2j)^2\mu^2}{2k\sigma^2}}
\end{aligned}$$

where $\Phi(\cdot)$ is the distribution function of $\mathcal{N}(0,1)$.

The above equation and (18), (22), (23) together lead to

$$\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| \leq \mu \sqrt{\frac{n}{m}} + \sigma \sqrt{\frac{2n}{\pi m}}.$$

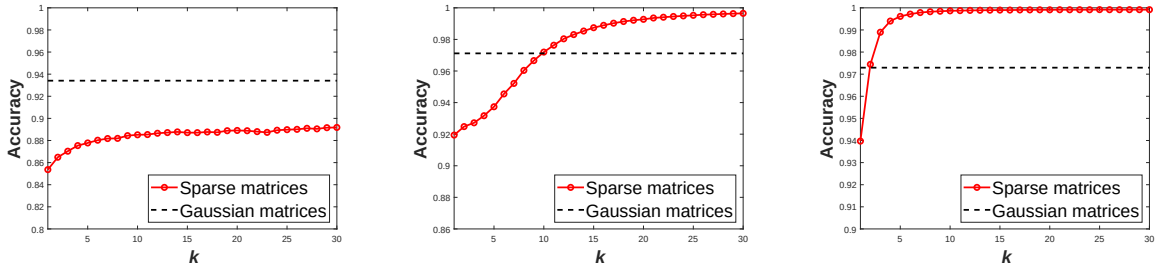
Next, we can derive the variance of $|\boldsymbol{r}^\top \boldsymbol{x}|$ as

$$\begin{aligned}
\mathrm{Var}(|\boldsymbol{r}^\top \boldsymbol{x}|) &= Var(\boldsymbol{r}^\top \boldsymbol{x}) - \left(\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| \right)^2 \\
&= \frac{n}{m}(\sigma^2 + 2q\mu^2) - \left(\mathbb{E}\|\boldsymbol{r}^\top \boldsymbol{x}\|_1 \right)^2.
\end{aligned}$$

Finally, the convergence of $\frac{\sqrt{m}}{\mu\sqrt{n}} \mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ shown in (12) and (13) can be derived in a similar way to the proof of P2 in Theorem 1. □

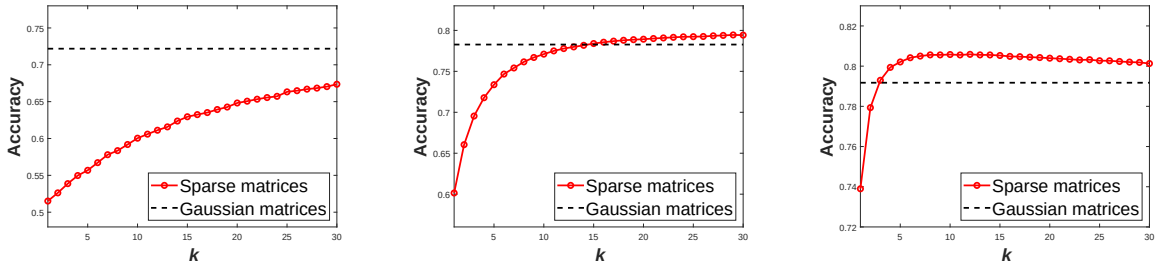
B Appendix

In Figs. 8–11, we test the SVM (with linear kernel) classification accuracy for the sparse ternary matrix with varying matrix sparsity k (and compression ratio m/n) on four different types of data. It can be seen that the performance changing trends of SVM against the varying matrix sparsity k are similar to the KNN performance as illustrated in the body of the paper, thus consistent with our theoretical analysis.



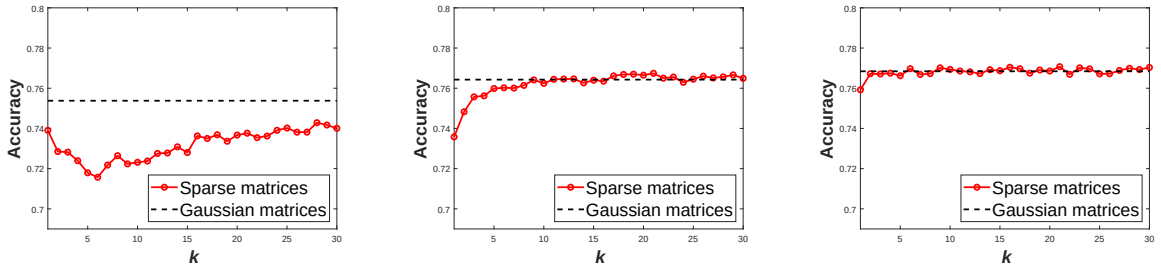
(a) $m/n = 1\%$ (b) $m/n = 10\%$ (c) $m/n = 50\%$

Figure 8: Classification accuracy of the sparse matrix-based and Gaussian matrix-based random projections for image data (YaleB, DCT features), with varying matrix sparsity $k \in [1, 30]$, three different projection ratios $m/n = 1\%$, 10% and 50%.



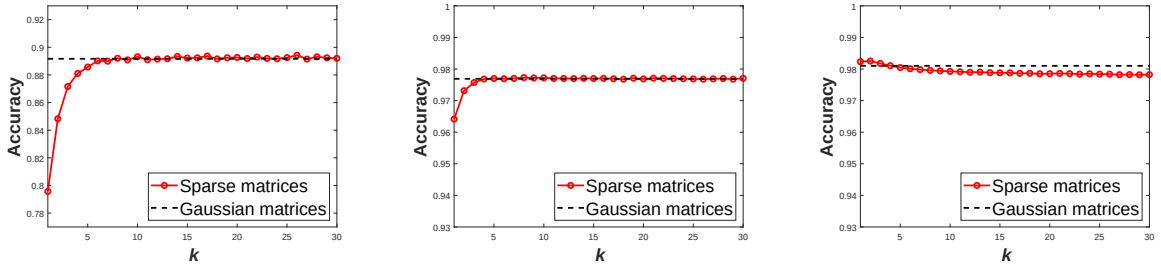
(a) $m/n = 1\%$ (b) $m/n = 10\%$ (c) $m/n = 50\%$

Figure 9: Classification accuracy of the sparse matrix-based and Gaussian matrix-based random projections for text data (Newsgroups), with varying matrix sparsity $k \in [1, 30]$, three different projection ratios $m/n = 1\%$, 10% and 50%.



(a) $m/n = 1\%$ (b) $m/n = 10\%$ (c) $m/n = 50\%$

Figure 10: Classification accuracy of the sparse matrix-based and Gaussian matrix-based random projections for gene data (AMLALL), with varying matrix sparsity $k \in [1, 30]$, three different projection ratios $m/n = 1\%$, 10% and 50%.



(a) $m/n = 1\%$ (b) $m/n = 10\%$ (c) $m/n = 50\%$

Figure 11: Classification accuracy of the sparse matrix-based and Gaussian matrix-based random projections for binary image data (MNIST, binarized pixels), with varying matrix sparsity $k \in [1, 30]$, three different projection ratios $m/n = 1\%$, 10% and 50%.