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 data dimension. For the problem, it is interesting to estimate the optimal sparsity of sparse matrices for classification, namely the minimum number of nonzero entries ± 1 that supports achieving the best classification performance. To achieve this, we analyze the impact of matrix sparsity on the ℓ_1 distance between projected data points. By principle component analysis, it is known 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 evaluate 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. Given the two data distributions, it is proved that the maximum ℓ_1 distance between projected data points could be approached, as the sparse matrix contains only one or at most about twenty nonzero entries per row, under the size $m \geq \mathcal{O}(\sqrt{n})$. Accordingly, the best classification performance should also be achieved under such conditions. This is confirmed with extensive experiments on different types of data, 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 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, and typical examples include Gaussian matrices (Dasgupta & Gupta, 1999) and sparse $\{0, \pm 1\}$-ternary matrices (shortly called sparse matrices hereafter) (Achlioptas, 2003). In practice, 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 interesting to minimize its complexity. For this purpose, we propose to estimate the optimal sparsity of sparse matrices for classification, namely, estimating the minimum number of nonzero entries ± 1 that allows the projected data to provide the 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 the sparsity of sparse matrices on the follow-on classification, since the classification

performance 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 means a worse classification performance, and in contrast the best classification performance can usually be achieved by very sparse matrices, such as the ones with only one nonzero entry per row. In the paper, we show that the intriguing performance can be explained 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).

For sparse matrices 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 estimate the optimal matrix sparsity that leads to the maximum ℓ_1 distance between projected data points, we need to first know and 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 one 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). As shown later, the use of the two general distributions ensures that our theoretical analysis and estimation are highly consistent with the experiments on real data.

Given the two data distributions, by varying the sparsity of sparse matrices, we estimate the *expected* ℓ_1 distance between projected data points and observe 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 difference vector between two original data contains a sufficient number of nonzero entries, namely holding a sufficiently dense distribution; 2) otherwise, the maximum distance tends to be reached by the matrices with at most about twenty nonzero entries per row. To summarize, the two results imply that the expected ℓ_1 distance between projected data points tends to reach its maximum value, as sparse matrices contain *only one or at most about twenty* nonzero entries per row. Moreover, we prove that the *expected* ℓ_1 distance can be approached with high probability by a single matrix sample, if the matrix has the size $m \times n$, $m \geq \mathcal{O}(\sqrt{n})$. This suggests that in practice a random matrix sample with the sparsity and size described above probably yields the maximum ℓ_1 distance between projected data points, and then results in the best classification performance. The performance is fully verified by conducting extensive experiments on a variety of real data, including the image, text, gene and binary quantization data. Overall, 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 estimate the optimal sparsity of sparse matrices for 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 tends to better account for the variation among original data and then yield better classification performance.
- Theoretically, we show that the optimal classification performance should be achieved by the sparse matrix with size $m \geq \mathcal{O}(\sqrt{n})$ and with only one or at most about twenty nonzero entries per row, if the original data holds the Gaussian mixture distribution or two-point distribution. The estimated optimal matrices exhibit very sparse structures, implying significant *savings* both in computation and storage.
- Empirically, we find that the theoretical estimations about the optimal matrix size and sparsity are highly consistent with the observations we have on the classification experiments with a variety of real data. The *consistency* between theory and practice benefits from the aforementioned two general distributions we have assumed for the original data, which can well approximate the distributions

of real data of different types (Torralba & Oliva, 2003; Weiss & Freeman, 2007; Wainwright & Simoncelli, 1999; Lam & Goodman, 2000).

2 Problem Formulation

In this section, we first model the distributions of sparse random matrices and original data, and then introduce the estimation model of the ℓ_1 distance between projected data points. 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. For ease of presentation, we defer all proofs to Appendix A.

2.1 The distribution of sparse matrices

The sparse random matrix $\boldsymbol{R} \in \{0, \pm\sqrt{\frac{n}{mk}}\}^{m\times n}$ we aim to study is described in Definition 1, which has the parameter k to count the number of nonzero entries per row, thus simply called k-sparse to distinguish between different matrix sparsity. Rather than simply defining $\boldsymbol{R} \in \{0, \pm 1\}^{m\times n}$, we introduce a scaling parameter $\sqrt{\frac{n}{mk}}$ to produce zero mean and unit variance for the matrix entries. This enables the matrix to hold the ℓ_2 distance preservation property, namely, keeping the expected ℓ_2 distance between original data points unchanged after random projection (Achlioptas, 2003). For easier computation, in practice the scaling parameter can be omitted and this will not change the relative distances between projected data points, thus not affecting the follow-on 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 with two-point distributions and Gaussian mixture distributions, we model the distribution of the pairwise *difference* between data points, which will be needed in the following ℓ_1 distance estimation.

2.2.1 Two-point distribution

Let $\boldsymbol{h}$, $\boldsymbol{h}' \in \{\mu_1, \mu_2\}^n$ be two high-dimensional data with each entry independently following a two-point distribution, where μ_1 and μ_2 are two arbitrary constants. Then the difference between the two data, expressed as $\boldsymbol{x} = \boldsymbol{h} - \boldsymbol{h}' = (x_1, x_2, \ldots, x_n)^\top$, 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

Suppose the two high-dimensional data $\boldsymbol{h}$, $\boldsymbol{h}' \in \mathbb{R}^n$ have their each entry independently following a Gaussian mixture distribution. Then the difference $\boldsymbol{x} = \boldsymbol{h} - \boldsymbol{h}'$ remains a Gaussian mixture (Andrews & Mallows, 1974), which allows each entry x_i 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 original data points $\boldsymbol{h}$, $\boldsymbol{h}' \in \mathbb{R}^n$ and k-sparse random matrix $\boldsymbol{R} \in \{0, \pm\sqrt{\frac{n}{mk}}\}^{m\times n}$, we aim to analyze the changing of the expected ℓ_1 distance between projected data points (namely $\mathbb{E}\|\boldsymbol{Rx}\|_1$, $\boldsymbol{x} = \boldsymbol{h}-\boldsymbol{h}'$) against the varying matrix sparsity k, so as to identify the sparsity k that maximizes the value of $\mathbb{E}\|\boldsymbol{Rx}\|_1$. By Definition 1, each row $\boldsymbol{r} \in \mathbb{R}^n$ of $\boldsymbol{R}$ follows an independent and identical distribution. Then we have $\mathbb{E}\|\boldsymbol{Rx}\|_1 = m\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}|$. The equivalence suggests that $\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}|$ will share the same changing trend with $\mathbb{E}\|\boldsymbol{Rx}\|_1$, when varying k. Then instead of $\mathbb{E}\|\boldsymbol{Rx}\|_1$, we propose to investigate $\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}|$ for ease of analysis.

3 The ℓ_1 Distance Estimation with Two-Point Distributed Data

In this section, we consider the original data with two-point distributions. Given such data, we estimate the value of $\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}|$ against varying matrix sparsity k, and identify the k values that lead to the maximum $\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}|$, where the difference $\boldsymbol{x} = \boldsymbol{h} - \boldsymbol{h}'$ between original data has i.i.d. entries $x_i \sim \mathcal{T}(\mu, p, q)$, as specified in equation 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

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

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 equation 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) $\lim\limits_{k\to\infty} \frac{\sqrt{m}}{\mu\sqrt{n}}\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}| = 2\sqrt{q/\pi}$.

In P1 and P2, we derive two results about the changing trend of $\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}|$ against the varying matrix sparsity k. By P1, $\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}|$ should achieve its maximum value at $k = 1$, as the probability p of x_i=0 is sufficiently small (≤ 0.188), namely, the difference $\boldsymbol{x}$ between data points exhibits sufficiently dense distributions; and by P2, $\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}|$ will converge to a stable value as k tends to infinity.

To more closely examine the changing trend, as detailed later, we conduct the numerical analysis and simulation for equation 4, and observe that: 1) the upper bound of p (shown in P1) that guarantees the maximum of $\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}|$ to be achieved at $k = 1$ can be further relaxed; 2) the maximum $\mathbb{E}|\boldsymbol{r}^\top\boldsymbol{x}|$ can be achieved at about $k = 20$, if the upper bound of p is violated; and 3) the speed of the convergence (shown in P2) is fast.

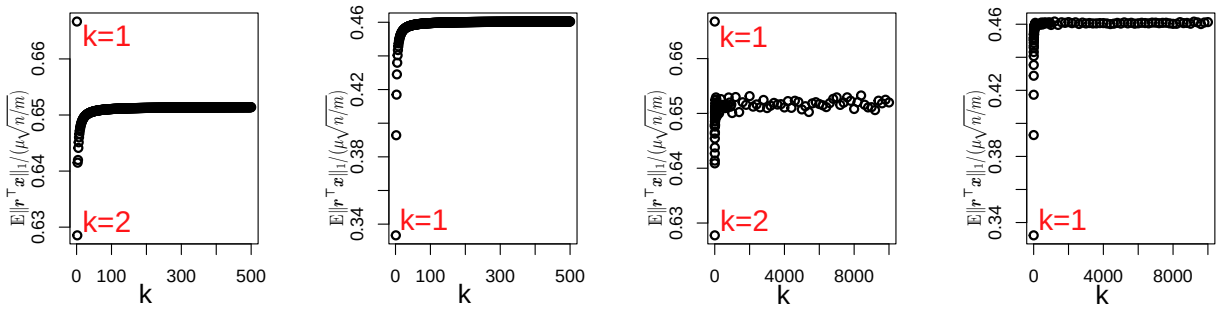


Figure 1: The value of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ calculated by equation 4 with $p = 1/3$ (a) and $p = 2/3$ (b), and estimated by numerical simulation with $p = 1/3$ (c) and $p = 2/3$ (d), provided $x_i \sim \mathcal{T}(\mu, p, q)$, $\mu = 1$.

3.2 Numerical analysis

By computing equation 4, we derive the values of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ against varying k in Figs. 1(a) and (b). Here we choose to study $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ instead of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$, mainly because both of them present the same changing trend against varying k, but the former has fewer parameters, i.e. only involving k and p. Recall that p is the probability of $x_i = 0$ in equation 1. By varying the continuous value of $p \in (0, 1)$ with fine steps and increasing the integer value of k from 1 to 500, we observe that:

(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 largest value at $k = 1$.

(P4) When $p > 1/3$, such as the case of $p = 2/3$ illustrated in Fig. 1(b), $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ inclines to first increase with k, and then quickly reach a stable level after about $k = 20$.

Compared to the theoretical analysis results P1 and P2, the numerical analysis results P3 and P4 are more positive. Specifically, P3 relaxes the upper bound of p from ≤ 0.188 to $\leq 1/3$. This suggests that $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ will achieve its maximum value at $k = 1$, when each entry x_i of $\boldsymbol{x}$ takes nonzero values with a probability greater than $2/3$, namely, the difference vector $\boldsymbol{x}$ between data points holds sufficiently dense distributions. Otherwise, by P4, $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ will reach its upper bound at about $k \geq 20$. *Therefore, we can reach the conclusion that* $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ *tends to reach its maximum value at* $k = 1$ *or at most about* $k = 20$*, for the original data with two-point distributions.*

3.3 Numerical simulation

To validate the numerical analysis results P3 and P4, we further investigate the changing trend of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ against varying k by numerical simulation. To do so, we randomly generate 10^6 pairs of $\boldsymbol{r}$ and $\boldsymbol{x}$ from their respective distributions, i.e. having $\boldsymbol{r} \in \{0, \pm\sqrt{\frac{n}{mk}}\}^n$ 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 $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/(\mu\sqrt{n/m})$ regarding k. Comparing the numerical results and simulation results provided in Fig. 1, namely (a) vs. (c) and (b) vs. (d), it is seen that the 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 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 can derive with a random matrix sample. To ensure the expected distance to be approached by actual samples, by the law of large numbers, the row size m of random matrix

$\boldsymbol{R}$ should be large enough. In Lemma 1, we derive a lower bound $m \geq \mathcal{O}(\sqrt{n})$ for the row size m by variance analysis.

Lemma 1. Let $z_1, z_2, \cdots, z_m$ be independent random variables with expectation equation 4 and variance equation 5, respectively. For all $\varepsilon, \delta > 0$, if $m \geq c \cdot \frac{\sqrt{n}}{\varepsilon\sqrt{\delta}}$, where $c = \sqrt{mVar(z_i)/n}$, then with probability at least $1 - \delta$, we have $|\bar{z} - \mathbb{E}\bar{z}| \leq \varepsilon$, where $\bar{z} = \frac{1}{m}\sum_{i=1}^{m} z_i$.

4 The ℓ_1 Distance Estimation with Gaussian Mixture Data

Similarly as in the previous section, we estimate the value of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ against the varying matrix sparsity k, and identify the value of k that leads to the maximum $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ for the original data drawn from Gaussian mixture distributions. In this case, the difference $\boldsymbol{x}$ between original data points has i.i.d. entries $x_i \sim \mathcal{M}(\mu, \sigma^2, p, q)$, as specified in equation 2.

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

$$\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| = 2\mu\sqrt{\frac{n}{mk}}T_1 + \sigma\sqrt{\frac{2n}{\pi m}}T_2 - 2\mu\sqrt{\frac{n}{mk}}T_3 \tag{6}$$

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

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

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

and

$$Var(|\boldsymbol{r}^\top \boldsymbol{x}|) = \frac{n}{m}(\sigma^2 + 2q\mu^2) - \left(\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|_1\right)^2 \tag{7}$$

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

$$\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| \leq \mu\sqrt{\frac{n}{m}} + \sigma\sqrt{\frac{2n}{\pi m}} \tag{8}$$

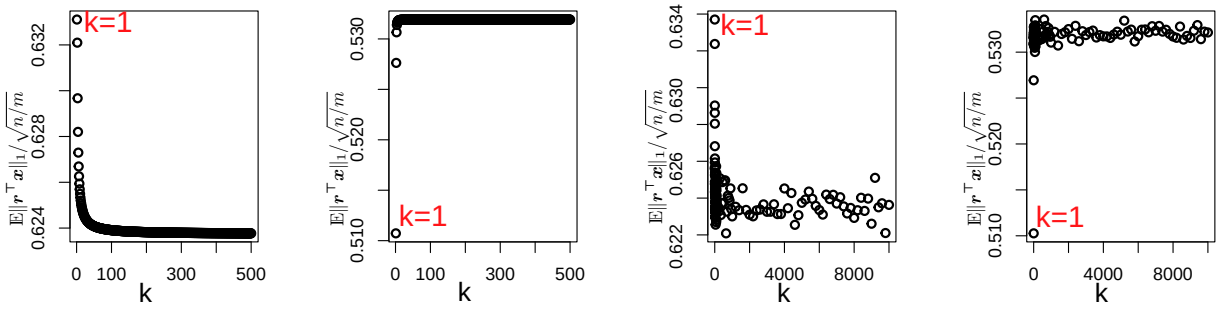
and

$$\lim_{k\to\infty} \frac{\sqrt{m}}{\mu\sqrt{n}}\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| = \sqrt{\frac{2}{\pi}(\sigma^2 + 2q\mu^2)}. \tag{9}$$

The theorem provides the expression of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ in equation 6 and proves its convergence in the limit of k in equation 9. By the following numerical analysis and simulation on equation 6, we will see that $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ is able to reach its maximum value at $k = 1$ or at most about $k = 20$.

4.2 Numerical analysis

In Figs. 2(a) and (b), by computing equation 6 we derive the value of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/\sqrt{n/m}$ across varying k. It is seen that $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/\sqrt{n/m}$ involves four parameters, k, p, μ, and σ. During computation, 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 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 $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/\sqrt{n/m}$ is not sensitive to σ/μ, but sensitive to p, the probability of each entry x_i of the data difference $\boldsymbol{x}$ taking zero value. More precisely, it is observed that



(a) (b) (c) (d)

Figure 2: The value of $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/\sqrt{n/m}$ calculated by equation 6 with $p = 1/2$ (a) and $p = 2/3$ (b), and estimated by numerical 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 $\sigma = 1/3$.

(P5) When $p \leq 1/2$, such as the case of $p = 1/2$ and $\sigma/\mu = 1/3$ shown in Fig. 2(a), $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/\sqrt{n/m}$ tends to first *decrease* with increasing k and then quickly reach a stable state after about $k = 20$.

(P6) When $p > 1/2$, such as the case of $p = 2/3$ and $\sigma/\mu = 1/3$ shown in Fig. 2(b), $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|/\sqrt{n/m}$ shows the trend of first *increasing* with k, and then reaching stable after about $k = 20$.

By P5, the maximum $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ can be achieved at $k = 1$, when each entry x_i of $\boldsymbol{x}$ takes nonzero values with a probability greater than $1/2$, or say, the data difference vector $\boldsymbol{x}$ has sufficiently dense distributions. Otherwise, by P6, the maximum $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ can be obtained at about $k \geq 20$. *To sum up, the two cases imply that* $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ *will reach its maximum value at* $k = 1$ *or at most about* $k = 20$*, for Gaussian mixture data with pairwise difference* $x_i \sim \mathcal{M}(\mu, \sigma^2, p, q)$.

It is noteworthy that the property is similar to the one we have derived in P3 and P4 for two-point distributed data with $x_i \sim \mathcal{T}(\mu, p, q)$. The similarity is not surprising since $x_i \sim \mathcal{T}(\mu, p, q)$ can be viewed as an extreme case of $x_i \sim \mathcal{M}(\mu, \sigma^2, p, q)$ with $\sigma \to 0$. Thanks to the good generalization capability of Gaussian mixture model, as will be seen in our experiments, the 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 approach the expected distance $\mathbb{E}\|\boldsymbol{R}^\top \boldsymbol{x}\|_1$ (equivalently $m\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$) derived with equation 6. The analysis is similar to Lemma 1, thus omitted here.

4.3 Numerical simulation

Now we are in the position to verify P5 and P6 by numerical simulation. Specifically, 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. 2(c) and (d), which have $p = 1/2$ and $2/3$, respectively. Comparing the numerical analysis and simulation results shown in Fig. 2, namely (a) vs. (c) and (b) vs. (d), it can be seen that both the two results are consistent with each other. The consistency validates the numerical analysis results P5 and P6.

5 Experiments

For the high-dimensional data with Gaussian-mixture distributions and two-point distributions, we have proved that the maximum ℓ_1 distance between their projections tends to be achieved by the sparse matrices with size $m \geq \mathcal{O}(\sqrt{n})$ and with exactly one or at most about twenty nonzero entries per row. Also, the best classification performance should be achieved by such matrices, in terms of the discrimination of large pairwise distances.

Considering many real data and their binary quantization can be approximately modeled respectively by Gaussian-mixture distributions and two-point distributions, in this section we aim to prove that the sparse matrix with the size and sparsity described above indeed performs best in practical classification problems.

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

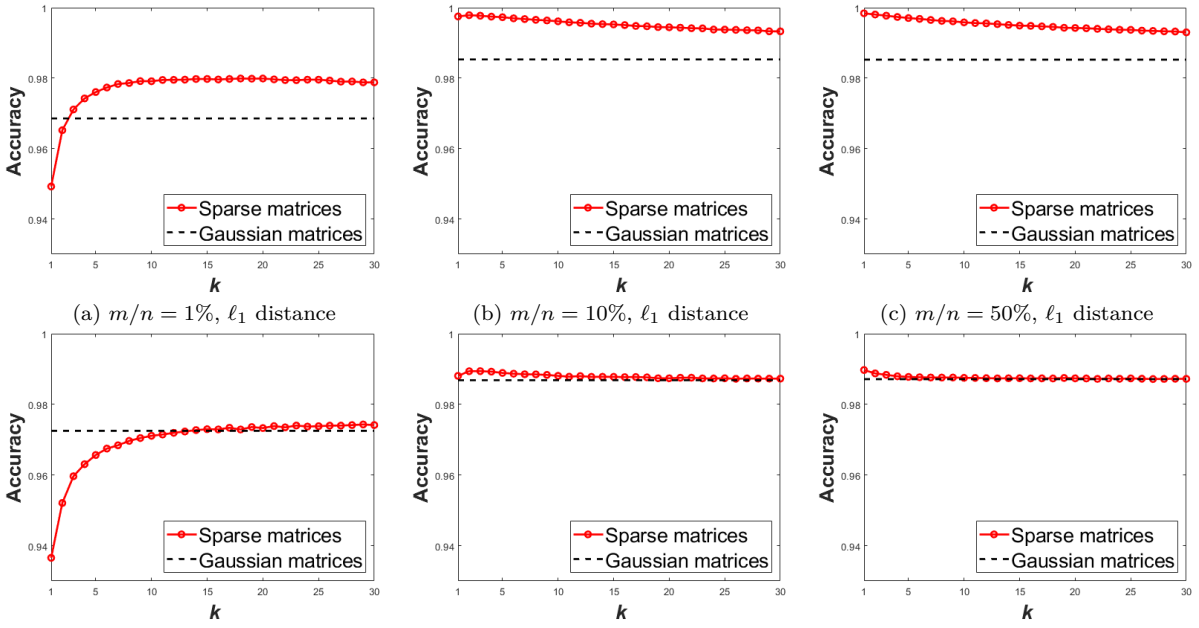
The random projection based classification is implemented by first multiplying original data with k-sparse random matrices and then classifying the resulting projections with a classifier. To faithfully reflect the impact of the varying data distance on classification, we adopt the simple nearest neighbor 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 optimal estimation could also be verified with other more sophisticated classifiers, like SVMs (Cortes & Vapnik, 1995), see Appendix B.

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 our optimal matrix is estimated with ℓ_1 distance, we also test and verify its performance advantage on the popular ℓ_2 distance.

5.3 Results

The classification results of four kinds of data are provided in Figs. 3–6, 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 $m/n = 1\%$.

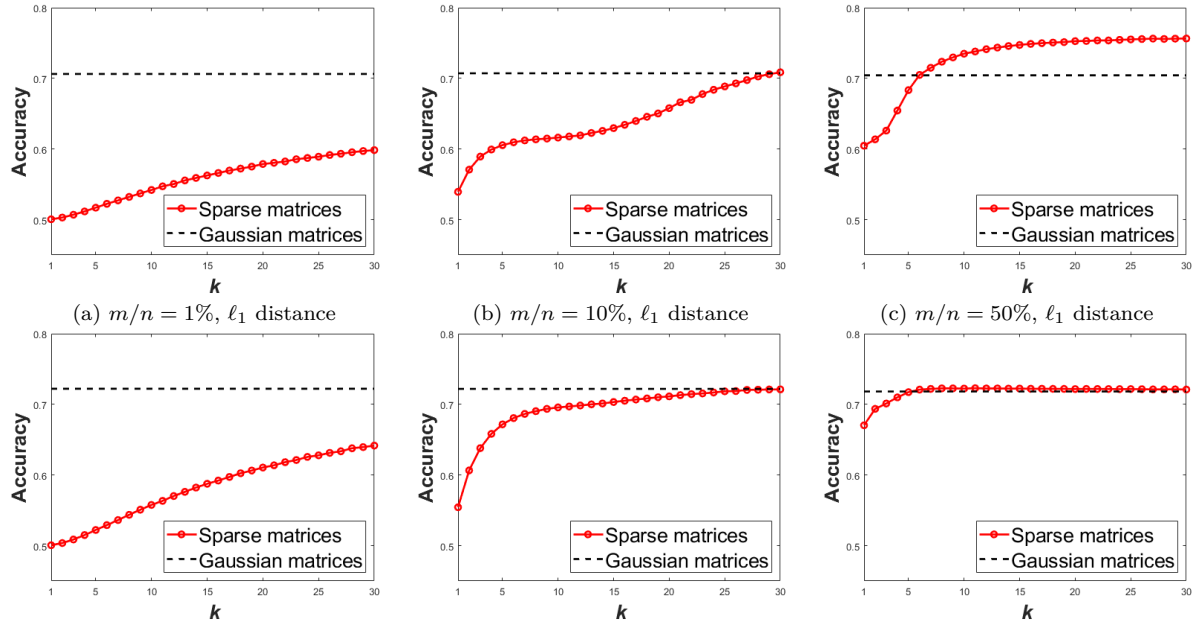
Let us first examine the case of satisfying $m \geq \mathcal{O}(\sqrt{n})$, namely the cases of $m/n = 10\%$ and 50% as shown in Figs. 3–6(b)(c). It is seen that the four kinds of data all achieve their best performance with relatively small matrix sparsity k ($\ll 30$), such as with $k = 1$ in Fig. 3(c) and $k = 6$ in Fig. 4(c). But in the case of $m/n = 1\%$ which violates the condition of $m \geq \mathcal{O}(\sqrt{n})$, as shown in Figs. 3–6(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 opposed to the cases of $m/n = 10\%$ and $m/n = 50\%$, it fails to get the decreasing performance trend and performs poorly at $k = 1$, as illustrated in Fig. 5. Overall, the experimental results on four different kinds of data all verify our theoretical estimation, that is the best classification performance can be achieved by the sparse matrices with only one or at most about twenty nonzero entries per row, under the size of $m \geq \mathcal{O}(\sqrt{n})$.



(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 3: 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.

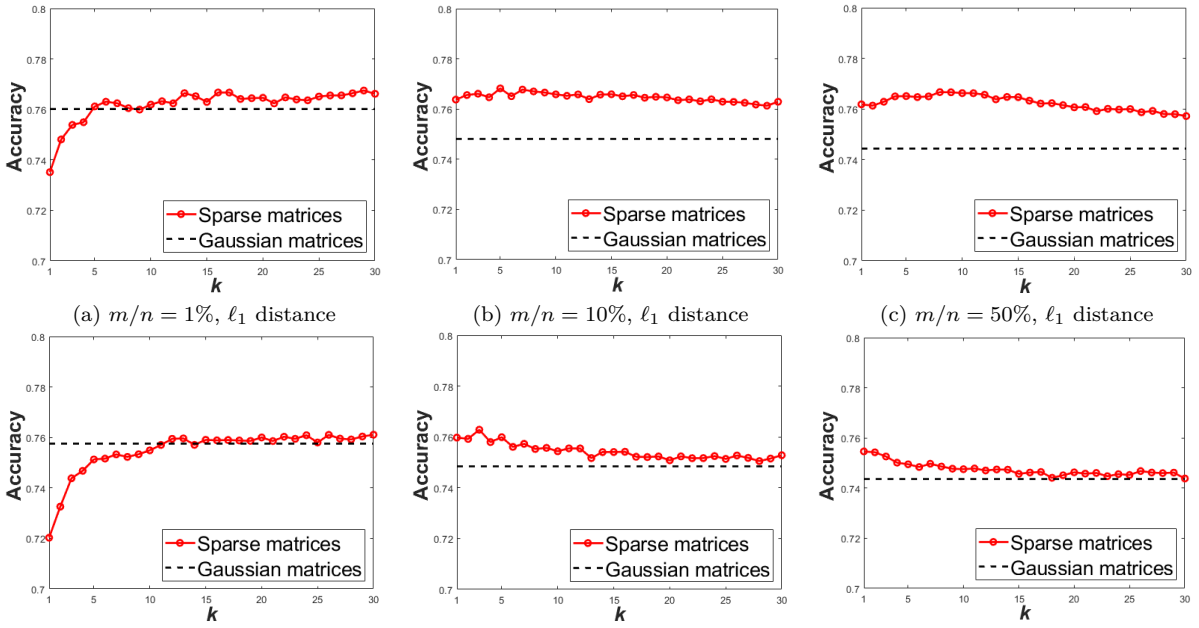


(a) $m/n = 1\%$, ℓ_1 distance (b) $m/n = 10\%$, ℓ_1 distance (c) $m/n = 50\%$, ℓ_1 distance

(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 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.

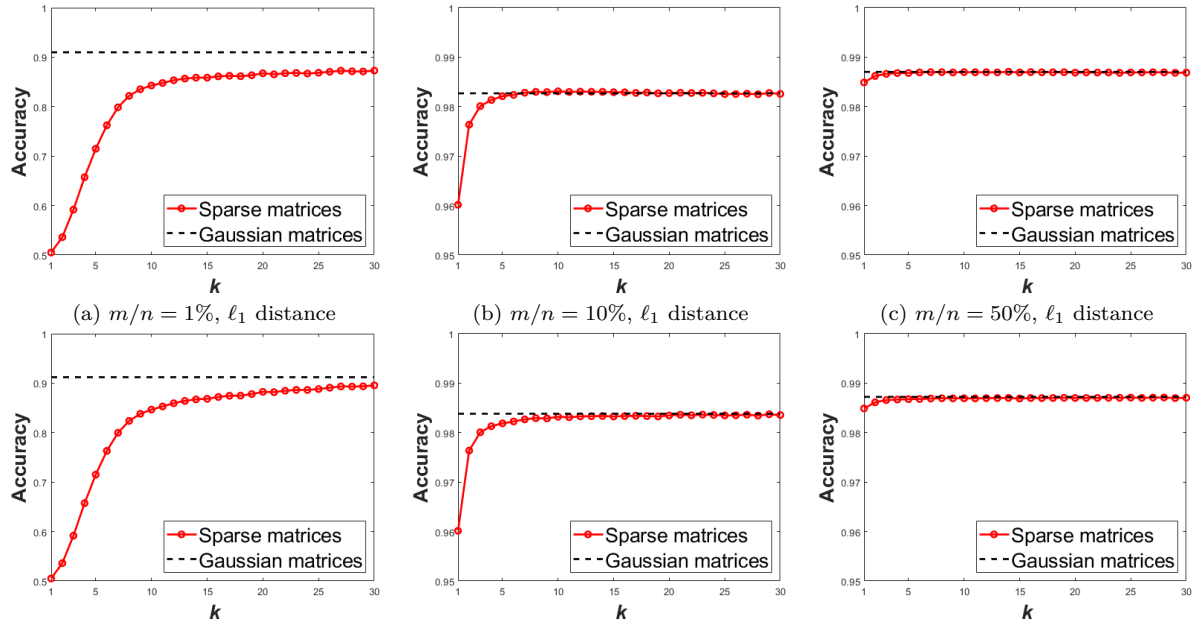
Besides the estimation of optimal matrix sparsity, the classification performance trend across varying matrix sparsity also consists with our estimation. More precisely, it can be seen from Figs. 3–6(b)(c) that the classifications of four datasets quickly converge to stable performance with the increasing matrix sparsity k. The difference between them mainly lies in the initial stage of the convergence. Specifically, as illustrated



(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 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.



(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 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.

in Figs. 3(b)(c) and 5(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 analysis result P5 depicted in Fig. 2(a). As for the curves on the other two datasets Newsgroups and MNIST, as shown in Figs. 4(b)(c)

and Figs. 6(b)(c), they both exhibit the trend of initially increasing with k, as predicted by the numerical analysis results P6 (illustrated in Fig. 2(b)) and P4 (illustrated in Fig. 1(b)).

Although our theoretical analysis is based on ℓ_1 distance, we can see that the classification with ℓ_2 distance exhibits similar performance trends, through comparing the results shown in the upper row versus the bottom row of Figs. 3–6. The similar performance of the two metrics in practical classification problems has been early observed and analyzed in (Gionis et al., 1999; Figiel et al., 1977). This suggests that the optimal sparse matrices we estimate with ℓ_1 distance also perform well in the classification with ℓ_2 distance, thus serving a wide range of applications. Moreover, the experiments show that sparse matrices often perform better than Gaussian matrices. This motivates us to employ sparse matrices instead of gaussian matrices, for its advantages both in complexity and accuracy.

6 Conclusion

For the sparse $\{0, \pm 1\}$-ternary matrix based random projection, we have demonstrated both in theory and practice that the best classification performance tends to be achieved by the sparse matrix with only one or at most about twenty nonzero entries per row, under the size of $m \geq \mathcal{O}(\sqrt{n})$. This implies that random projection can be implemented with extremely sparse matrices, which have significant advantages in storage and computation compared to the popularly used Gaussian matrices. Impressively, our theoretical estimation exhibits high consistency with the experimental evaluation on real data of different types, such as the image, text, gene and binary quantization data. The high consistency between theory and practice can be attributed to the good generalization of the two data distributions we have assumed for statistical analysis, i.e. the Gaussian mixture distribution and two-point distribution. Besides the major contribution described above, there are three other important results worth mentioning.

First, the optimal sparse matrix we estimate with ℓ_1 distance also performs well for the ℓ_2 distance-based classification. The generalization can be attributed to the closeness of the two metrics, which has been found and analyzed in early research (Gionis et al., 1999; Figiel et al., 1977). Overall, this is good news in terms of the wide applications of the two metrics (Philbin et al., 2008).

Second, experiments show that our sparse matrices tend to provide higher classification accuracy than the popularly used Gaussian matrices. This encourages us to employ sparse matrices instead of gaussian matrices, for improvements both in complexity and accuracy.

Third, our optimal matrix sparsity estimation 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 optimal matrix sparsity estimation for random projection-based classification can be viewed as a layerwise analysis of deep ternary networks. The very sparse ternary matrices we derive for optimal classification partly explains the performance advantage 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.

James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *2008 IEEE conference on computer vision and pattern recognition*, pp. 1–8. IEEE, 2008.

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.

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. doi: 10.1109/CVPR.1991.139758.

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 equation 4, equation 5, P1 and P2.

Proofs of equation 4 and equation 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

$$\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| = \mu\sqrt{\frac{n}{mk}} \sum_{i=0}^{k} C_k^i p^i q^{k-i} \sum_{j=0}^{k-i} C_{k-i}^j |k-i-2j|, \tag{10}$$

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

$$\sum_{j=0}^{k-i} (C_{k-i}^j |k-i-2j|) = 2\left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil}, \tag{11}$$

where $\lceil \alpha \rceil = \min\{\beta : \beta \geq \alpha, \beta \in \mathbb{Z}\}$. Combining (10) and (11), we can obtain equation 4.

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

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

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 equation 4, it is easy to derive that

$$\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|_{k=1} = 2q\mu\sqrt{\frac{n}{m}}. \tag{12}$$

Then, let us see the case of computing $\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|_{k>1}$. By equation 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

$$\begin{aligned} &\frac{2}{\sqrt{k}} C_k^i p^i q^{k-i} \left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil} \\ &\leq \frac{1}{\sqrt{k}} C_k^i p^i q^{k-i} (k-i) 2^{k-i} \sqrt{\frac{2}{(k-i)\pi}} \\ &\leq \sqrt{\frac{2}{\pi}} C_k^i p^i (2q)^{k-i}, \end{aligned} \tag{13}$$

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

$$\begin{aligned} &\frac{2}{\sqrt{k}} C_k^i p^i q^{k-i} \left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil} \\ &\leq \frac{1}{\sqrt{k}} C_k^i p^i q^{k-i} (k-i) 2^{k-i} \sqrt{\frac{2}{(k-i-1)\pi}} \\ &= \sqrt{\frac{2}{\pi}} C_k^i p^i (2q)^{k-i} \frac{k-i}{\sqrt{k(k-i-1)}} \end{aligned} \tag{14}$$

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,

$$\begin{aligned} &\frac{2}{\sqrt{k}} C_k^i p^i q^{k-i} \left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil} \\ \leq &\begin{cases} \sqrt{\frac{2}{\pi}} C_k^i p^i (2q)^{k-i}, & k \geq 5, i \geq 2, \\ \frac{2}{\sqrt{k}} C_k^i p^i q^{k-i} (k-i) C_{k-i-1}^{\frac{k-i-1}{2}}, & \text{otherwise.} \end{cases} \end{aligned} \tag{15}$$

According to the results equation 13 and equation 15 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 equation 4, we have

$$\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| = \begin{cases} \frac{\mu\sqrt{n}}{\sqrt{2m}}(4q^2 + 4pq), & k = 2, \\ \frac{\mu\sqrt{n}}{\sqrt{3m}}(12q^3 + 12pq^2 + 6p^2 q), & k = 3, \\ \frac{\mu\sqrt{n}}{\sqrt{m}}(12q^4 + 24pq^3 + 12p^2 q^2 + 4p^3 q), & k = 4, \end{cases} \tag{16}$$

and for the case of $k \geq 5$, with equation 13 and equation 15, we have

$$\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| \leq \mu\sqrt{\frac{2n}{\pi m}} + \mu\sqrt{\frac{n}{m}}(2q)^5 \left(\frac{3\sqrt{5}}{8} - \sqrt{\frac{2}{\pi}} \right). \tag{17}$$

By equation 12, equation 16 and equation 17, 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 equation 17 by considering two cases of k, being even or odd.

Case 1: Suppose $k \geq 5$ and k is even. Combining equation 13 and equation 15, we have

$$\begin{aligned} \mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| \leq &\mu\sqrt{\frac{n}{m}} C_k^1 p (2q)^{k-1} \left(\frac{\sqrt{k}}{2^{k-1}} C_{k-1}^{\frac{k}{2}-1} - \sqrt{\frac{2}{\pi}} \right) \\ &+ \mu\sqrt{\frac{2n}{\pi m}} \sum_{i=0}^{k} C_k^i p^i (2q)^{k-i}. \end{aligned} \tag{18}$$

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

$$h_1(k) = \frac{\sqrt{k}}{2^{k-1}} C_{k-1}^{\frac{k}{2}-1} \leq \lim_{k\to\infty} h_1(k) = \sqrt{\frac{2}{\pi}}. \tag{19}$$

Then, it follows from (18) and (19) that

$$\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| \leq \mu\sqrt{\frac{2n}{\pi m}}. \tag{20}$$

Case 2: Suppose $k \geq 5$ and k is odd. Combining (13) and (15), we have

$$\begin{aligned}\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}| \leq & \mu\sqrt{\frac{n}{m}} C_k^0 (2q)^k \left(\frac{\sqrt{k}}{2^{k-1}} C_{k-1}^{\frac{k-1}{2}} - \sqrt{\frac{2}{\pi}}\right) \\ & + \mu\sqrt{\frac{2n}{\pi m}} \sum_{i=0}^{k} C_k^i p^i (2q)^{k-i}.\end{aligned} \tag{21}$$

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

$$h_2(k) = \frac{\sqrt{k}}{2^{k-1}} C_{k-1}^{\frac{k-1}{2}} \leq h_2(5) = \frac{\sqrt{5}}{2^4} C_4^2. \tag{22}$$

Then, it follows from (21) and (22) 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

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

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

$$\frac{1}{\sqrt{k}} \sum_{i=1}^{k} z_i \rightsquigarrow Z, \tag{23}$$

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

Then based on equation 17, 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

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

□

A.2 Proof of Lemma 1

Proof. With the definition of z, we have $\mathbb{E}z = \sigma\sqrt{\frac{2n}{\pi m}}$ and $Var(z) = \frac{n}{m^2}(1-\frac{2}{\pi})\sigma^2$. Then using the Chebyshev's Inequality,

$$\mathbf{Pr}\{|z-\mathbb{E}z| > \varepsilon\} \le \frac{Var(z)}{\varepsilon^2} \le \delta.$$

Therefore, $m \ge c\cdot\frac{\sqrt{n}}{\varepsilon\sqrt{\delta}}$, where $c = \sigma\sqrt{1-2/\pi}$. □

A.3 Proof of Theorem 2

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

$$\mathbb{E}|z| = \sqrt{\frac{2}{\pi}}\sigma e^{-\frac{\mu^2}{2\sigma^2}} + \mu\left(1-2\Phi\left(-\frac{\mu}{\sigma}\right)\right) \tag{24}$$

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

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

Then, by equation 24 we can derive that

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

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

The above equation and equation 12, equation 16, equation 17 together lead to

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

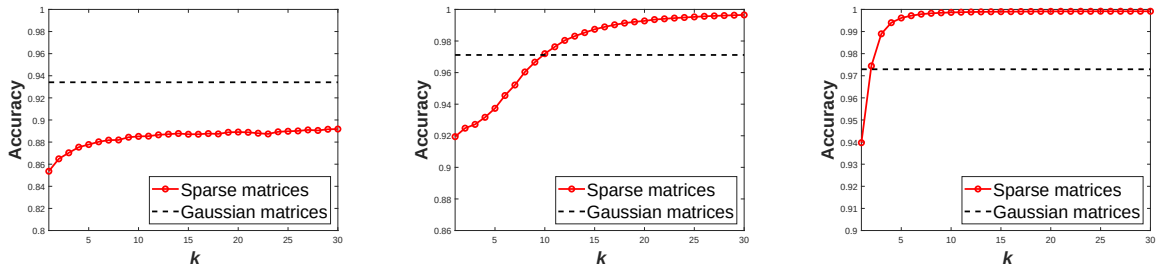
Next, we can derive the variance of $|\boldsymbol{r}^\top \boldsymbol{x}|$ as

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

Finally, the convergence of $\frac{\sqrt{m}}{\mu\sqrt{n}}\mathbb{E}|\boldsymbol{r}^\top \boldsymbol{x}|$ shown in equation 9 can be derived by the same method of proving P2. □

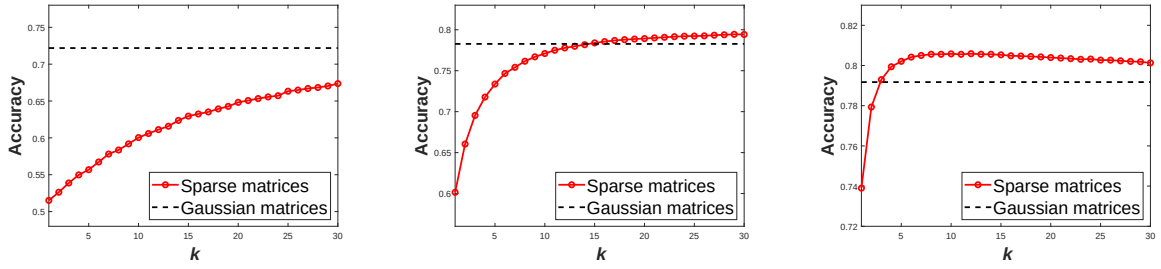
B Appendix

In Figs. 7–10, 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.



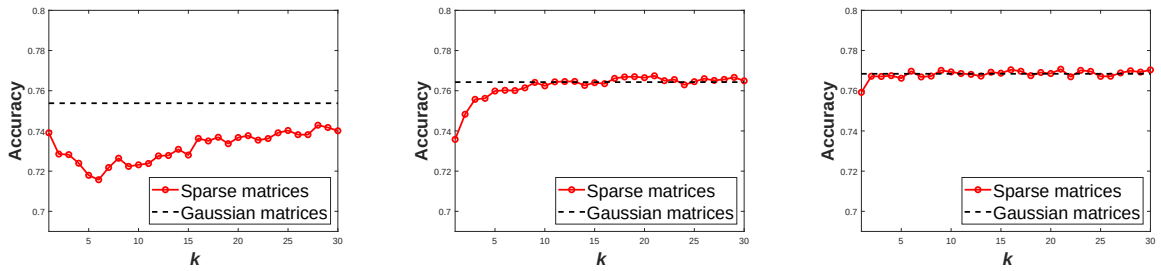
(a) $m/n = 1\%$ (b) $m/n = 10\%$ (c) $m/n = 50\%$

Figure 7: 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%.



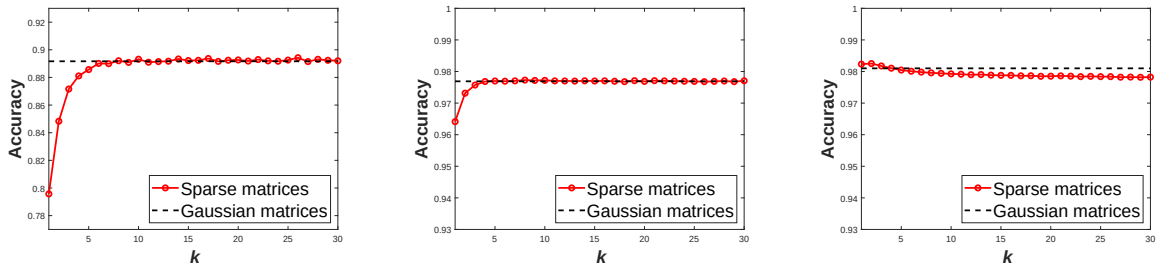
(a) $m/n = 1\%$ (b) $m/n = 10\%$ (c) $m/n = 50\%$

Figure 8: 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 9: 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 10: 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%.