

THE SPARSE MATRIX-BASED RANDOM PROJECTION: EXPLORING OPTIMAL 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, a technique extensively applied in diverse classification tasks for dimensionality reduction and as a foundational model for each layer in the popular deep ternary networks. For these sparse matrices, determining the optimal sparsity level, namely the minimum number of nonzero entries ± 1 needed to achieve the optimal or near-optimal classification performance, remains an unresolved challenge. To investigate the impact of matrix sparsity on classification, we propose to analyze the mean absolute deviation (MAD) of projected data points, which quantifies their dispersion. Statistically, a higher degree of dispersion is expected to improve classification performance by capturing more intrinsic variations in the original data. Given that the MAD value depends not only on the sparsity level of random matrices but also on the distribution of the original data, we evaluate two representative data distributions for generality: the Gaussian mixture distribution, widely used to model complex real-world data; and the two-point distribution, available for modeling discretized data. Our analysis reveals that sparse matrices with only *one* or *a few* nonzero entries per row can achieve MAD values comparable to, or even exceed, those of denser matrices, provided the matrix size satisfies $m \geq \mathcal{O}(\sqrt{n})$, where m and n denote the projected and original dimensions, respectively. These extremely sparse matrix structures imply significant computational savings. This finding is further validated through classification experiments on diverse real-world datasets, including images, text, gene data, and binary-quantized data, demonstrating its broad applicability.

1 INTRODUCTION

Random projection is a fundamental, unsupervised dimensionality reduction technique that projects high-dimensional data onto low-dimensional subspaces using random matrices (Johnson & Lindenstrauss, 1984). This projection method can approximately maintain the pairwise ℓ_2 distance between original data points, preserving the structure of original data and making it suitable for classification tasks (Bingham & Mannila, 2001; Fradkin & Madigan, 2003; Wright et al., 2009). To ensure the preservation of ℓ_2 distances, random projection matrices must adhere to specific distributions, such as Gaussian matrices (Dasgupta & Gupta, 1999) and sparse $\{0, \pm 1\}$ -ternary matrices (hereinafter referred to as sparse matrices) (Achlioptas, 2003). In practical scenarios, sparse matrices are favored due to their significantly lower storage and computational complexity. Given that random projection is extensively used in computationally intensive, large-scale classification tasks (Gionis et al., 1999), and can even model each layer of deep networks (Giryes et al., 2016), it is highly desirable to investigate the optimal sparsity of sparse matrices, namely the minimum number of nonzero entries ± 1 required for the projected data to achieve the optimal or near-optimal classification performance. To the best of our knowledge, this problem has not been previously investigated.

Existing random projection research has primarily focused on characterizing the distribution of random matrices that effectively preserve pairwise distances. Specifically, these matrices are designed to maintain the expected pairwise distances between original data points after projection, while keeping the variance relatively small (Dasgupta & Gupta, 1999; Achlioptas, 2003). For sparse matrices with appropriately scaled entries, distance preservation has been validated in the ℓ_2 norm (Achlioptas, 2003; Li et al., 2006), but fails in the ℓ_1 norm (Brinkman & Charikar, 2003; Li, 2007). While the ℓ_2 distance preservation property supports classification tasks, it does not lend itself well

to analyzing the influence of matrix sparsity (i.e., the number of nonzero entries) on subsequent classification accuracy. This is because classification accuracy depends on the discriminative power among projected points rather than strict preservation of original data structure. For example, ℓ_2 distance preservation degrades as the matrix becomes sparser, namely containing fewer nonzero entries ± 1 (Li et al., 2006). However, empirical studies reveal that sparser matrices does not necessarily lead to worse classification performance; in fact, extremely sparse matrices, such as those with only one or a few dozen nonzero entries per row, often achieve comparable or superior classification accuracy compared to denser alternatives. This counterintuitive phenomenon remains theoretically unexplained.

In the paper, we demonstrate that the performance advantages of extremely sparse matrices can be explained by analyzing the dispersion of projected data points. Statistically, a higher degree of dispersion should lead to improved classification performance, as it captures more intrinsic variations in the original data (Jolliffe & Cadima, 2016). Generally, dispersion is quantified using variances, as seen in principal component analysis (PCA) (Jolliffe, 2002), primarily due to the computational and analytical convenience of variances. However, variance-based dispersion analysis is ideally suited for Gaussian data and sensitive to noise and outliers (Hubert et al., 2016). Such limitations are common in real-world applications. In these cases, the mean absolute deviation (MAD) (Yager & Alajlan, 2014) offers a more robust alternative for quantifying dispersion (Deng et al., 2007; McCoy & Tropp, 2011). Consequently, we adopt MAD in our study to identify the matrix sparsity that maximizes dispersion, thereby enhancing classification performance.

The MAD value depends not only on the sparsity level of random matrices, but also on the distribution of the original data. For the sake of generality, we consider two representative data distributions: the Gaussian mixture distribution, which has been extensively used to model natural data distributions (Torralba & Oliva, 2003; Weiss & Freeman, 2007) and their feature transformations (Wainwright & Simoncelli, 1999; Lam & Goodman, 2000); and the two-point distribution, which is suitable for modelling discretized data, a scenario gaining increasing relevance with the emergence of large-scale data and models (Gionis et al., 1999; Hubara et al., 2016; Yang et al., 2019). Benefiting from the two fundamental distributions, as demonstrated later, our theoretical analysis results can be generalized to a broad range of real-world data scenarios.

By analyzing the MAD of projected data points across varying levels of matrix sparsity, we identify two major results. First, sparse matrices with exactly one nonzero entry per row tends to achieve best classification performance, as the original data exhibit sufficiently high discrimination between individuals. Second, as matrix sparsity increases, classification accuracy converges asymptotically to a stable plateau; notably, this convergence occurs even with moderate sparsity levels, such as the level of containing only a few dozen nonzero entries per row. Both findings hold under the condition $m \geq \mathcal{O}(\sqrt{n})$, where m and n denote the projected and original dimensions, respectively. Collectively, the above two results demonstrate that extremely sparse matrices with only *one* or *a few* nonzero entries per row perform comparably or superiorly to denser alternatives in classification tasks. This performance advantage is empirically validated across various real-world datasets, including images, texts, gene expression data, and binary-quantized data. Crucially, our findings imply that the computational complexity of sparse matrices-based random projection can be drastically reduced, while maintaining or even enhancing downstream classification accuracy. This breakthrough represents a significant efficiency improvement for random projection-based models, particularly for high-dimensional data applications.

2 PROBLEM FORMULATION

Consider a random projection $\mathbf{z} = \mathbf{R}\mathbf{h}$, where $\mathbf{h} \in \mathbb{R}^n$ denotes an original data and $\mathbf{R} \in \{0, \pm 1\}^{m \times n}$ is a sparse random matrix. In the paper, we aim to estimate the optimal sparsity level of $\mathbf{R}$, namely the minimum number of nonzero entries ± 1 , that enables maximizing the mean absolute deviation (MAD) of projected data points $\mathbf{z} \in \mathbb{R}^m$, formally defined as $\text{MAD}(\mathbf{z}) = \mathbb{E} \|\mathbf{z} - \mathbb{E}\mathbf{z}\|_1$. As discussed before, a larger $\text{MAD}(\mathbf{z})$ is expected to yield better classification performance. Since $\text{MAD}(\mathbf{z})$ involves both the distributions of random matrices $\mathbf{R}$ and original data $\mathbf{h}$, in the sequence we first specify their probabilistic models before detailing the MAD estimation procedure.

Notation. Throughout the paper, we typically denote a matrix by a bold upper-case letter $\mathbf{R} \in \mathbb{R}^{m \times n}$, a vector by a bold lower-case letter $\mathbf{r} = (r_1, r_2, \dots, r_n)^\top \in \mathbb{R}^n$, and a scalar by a lower-case

letter r_i or r . Sometimes, we use the bold letter $\mathbf{r}_i \in \mathbb{R}^n$ to denote the i -th row of $\mathbf{R} \in \mathbb{R}^{m \times n}$. For ease of presentation, we defer all proofs to the appendix.

2.1 THE DISTRIBUTION OF SPARSE MATRICES

The sparse random matrix $\mathbf{R}$ 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 $\mathbf{R} \in \{0, \pm 1\}^{m \times n}$, 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 downstream classification performance.

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

- Each row $\mathbf{r} \in \{0, \pm \sqrt{\frac{n}{mk}}\}^n$ 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 data $\mathbf{h} = (h_1, h_2, \dots, h_n)^\top$, as discussed previously, we investigate two representative distributions: two-point distributions and Gaussian mixture distributions, which are detailed below. When analyzing $\text{MAD}(\mathbf{z})$, the difference $\mathbf{x}$ between two arbitrary data points $\mathbf{h}$ and $\mathbf{h}'$, namely $\mathbf{x} = \mathbf{h} - \mathbf{h}' = (x_1, x_2, \dots, x_n)^\top$, will be involved. Then the distribution of the difference vector $\mathbf{x}$ is also introduced.

2.2.1 TWO-POINT MIXTURE DISTRIBUTION

The two-point distribution is modeled as $\mathbf{h} \in \{\mu_1, \mu_2\}^n$, $\mu_1 \neq \mu_2$, with each entry $h_i \in \{\mu_1, \mu_2\}$ independently and identically distributed. As mentioned before, this distribution widely exists in quantization tasks, such as the quantization of deep networks (Hubara et al., 2016).

For the difference vector $\mathbf{x}$ between two data points $\mathbf{h}$ and $\mathbf{h}'$, its each entry x_i will follow a ternary discrete distribution

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

with the probability mass function $t \in \{-\mu, 0, \mu\}$ under the probabilities $\{q, p, q\}$, where $\mu = \mu_1 - \mu_2$ and $p + 2q = 1$. Note that a smaller p value (equivalently, a larger q value) indicates a greater number of nonzero entries in x_i , suggesting a higher degree of discrimination between h_i and h'_i .

2.2.2 GAUSSIAN MIXTURE DISTRIBUTION

For the Gaussian mixture distribution, we assume two Gaussian components for each entry of the original data $\mathbf{h}$: $h_i \sim \sum_{j=1}^2 \omega_j \mathcal{N}(\mu_j, \sigma^2)$, where $\mu_1 \neq \mu_2$, $\omega_1 + \omega_2 = 1$, and $\omega_j \geq 0$. With the two components, we can model the distribution of feature values in each dimension as having two underlying states: "strong" and "weak". Each Gaussian component would represent one of these states, with its mean and variance characterizing the typical values and spread for that intensity level. This modeling method is logically sound and applicable to various areas (Rowe & Blake, 1995; Wang et al., 2014; Xing et al., 2001; Mulè et al., 2022; Zapevalov & Knyazkov, 2023), striking a balance between model complexity and generalization ability.

The difference vector $\mathbf{x}$ between two data points $\mathbf{h}$ and $\mathbf{h}'$ has each entry x_i satisfying a three-component Gaussian mixture distribution

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

with the probability density function

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

where $f_{\mathcal{N}}(t; \mu, \sigma^2)$ denotes the density function of $t \sim \mathcal{N}(\mu, \sigma^2)$, and p and q represent the mixture weights of three components, with $p + 2q = 1$, and $p, q \geq 0$. As in the two-point distribution equation 1, a smaller p value indicates a higher degree of discrimination between h_i and h'_i .

2.3 THE MAD ESTIMATION FOR PROJECTED DATA

As noted at the beginning of this section, our objective is to estimate the minimum sparsity k of the matrix $\mathbf{R} \in \{0, \pm\sqrt{\frac{n}{mk}}\}^{m \times n}$ that maximizes the $\text{MAD}(\mathbf{z})$ of the projected data $\mathbf{z} = \mathbf{R}\mathbf{h}$. We now show that analyzing how $\text{MAD}(\mathbf{z})$ varies with k can be simplified to studying $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$, where $\mathbf{x} = \mathbf{h} - \mathbf{h}'$ and $\mathbf{r}$ is a row of $\mathbf{R}$.

By Property 1, when the original data $\mathbf{h}$ follow the Gaussian mixture distribution described above, the projected data $\mathbf{z}$ remain Gaussian. Moreover, $\text{MAD}(\mathbf{z})$ is a positive constant multiple of $\mathbb{E}|\mathbf{z} - \mathbf{z}'|_1$, where $\mathbb{E}|\mathbf{z} - \mathbf{z}'|_1 = \mathbb{E}|\mathbf{R}\mathbf{x}|_1$ and $\mathbf{x} = \mathbf{h} - \mathbf{h}'$. This implies that $\mathbb{E}|\mathbf{R}\mathbf{x}|_1$ varies in the same way as $\text{MAD}(\mathbf{z})$ when the matrix sparsity k changes. In particular, a larger value of $\mathbb{E}|\mathbf{R}\mathbf{x}|_1$ corresponds to a larger $\text{MAD}(\mathbf{z})$. This relationship also holds approximately for original data $\mathbf{h}$ drawn from other distributions, since by the Central Limit Theorem, the projected data $\mathbf{z} \in \mathbb{R}^m$ can be approximated by a Gaussian distribution. Therefore, rather than analyzing $\text{MAD}(\mathbf{z})$ directly, we can examine how $\mathbb{E}|\mathbf{R}\mathbf{x}|_1$ changes with k .

Note that the analysis of $\mathbb{E}|\mathbf{R}\mathbf{x}|_1$ can be further simplified by focusing on $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$, where $\mathbf{r} \in \mathbb{R}^n$ is a row of $\mathbf{R}$, as the latter exhibits the same trend as the former, with respect to variations in k . This equivalence arises from the fact that $\mathbb{E}|\mathbf{R}\mathbf{x}|_1 = m\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$, since each row $\mathbf{r}$ of $\mathbf{R}$ is independently and identically distributed by Definition 1. Therefore, $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ is adopted as our metric to characterize the relationship between $\text{MAD}(\mathbf{z})$ and matrix sparsity k . A larger $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ value corresponds to a higher $\text{MAD}(\mathbf{z})$, then yielding improved classification performance.

Property 1. Consider $\mathbf{z} = \mathbf{R}\mathbf{h}$, where $\mathbf{h} \in \mathbb{R}^n$ follows the Gaussian mixture distribution described in Section 2.2.2 and $\mathbf{R} \in \{0, \pm 1\}^{m \times n}$ is distributed as specified in Definition 1, with its scale parameter omitted for brevity. Considering Gaussian distributions are closed under linear transformations, the i.i.d. entries z_i of $\mathbf{z}$ still holds a Gaussian distribution: $z_i \sim \mathcal{N}(\mu, \sigma^2)$. For this distribution of $\mathbf{z}$, we have $\mathbb{E}|\mathbf{z} - \mathbb{E}\mathbf{z}|_1 = \frac{1}{\sqrt{2}}\mathbb{E}|\mathbf{z} - \mathbf{z}'|_1$, where $\mathbf{z}'$ is an independent copy of $\mathbf{z}$.

3 THEORETICAL RESULTS

3.1 TWO-POINT DATA

As analyzed in Section 2.3, the trend of $\text{MAD}(\mathbf{z})$ against varying matrix sparsity k can be estimated with $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$. A larger $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ corresponds to a higher $\text{MAD}(\mathbf{z})$, thereby achieving improved classification performance. Based on this relationship, we analyze how $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ changes with respect to k , assuming that the original data vectors $\mathbf{h}$ and $\mathbf{h}'$ are drawn from a two-point distribution, such that their difference vector $\mathbf{x} = \mathbf{h} - \mathbf{h}'$ has i.i.d. entries $x_i \sim \mathcal{T}(\mu, p, q)$ as defined in equation 1.

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

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

and

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

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, $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ satisfies the following two properties:

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

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

$$\lim_{k \rightarrow \infty} \frac{\sqrt{m}}{\mu\sqrt{n}} \mathbb{E}|\mathbf{r}^\top \mathbf{x}| = 2\sqrt{q/\pi}, \quad (6)$$

which has the convergence error upper-bounded by

$$\left| \frac{\sqrt{m}}{\mu\sqrt{n}} \mathbb{E}|\mathbf{r}^\top \mathbf{x}| - 2\sqrt{q/\pi} \right| \leq \frac{\sqrt{\pi} + \sqrt{2}}{\sqrt{\pi k}}, \quad (7)$$

for finite k values.

Remarks of Theorem 1: In P1 and P2, we demonstrate two distinct trends of $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ against varying matrix sparsity k , indicating two corresponding changes in classification performance, which are elaborated as follows.

- P1 indicates that the best classification performance can be achieved by very sparse random matrices with sparsity $k = 1$, when the discrimination between data points is sufficiently high. This can be explained as follows. By P1, $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ will achieve its maximum value at $k = 1$, if the probability p of $x_i = 0$ is sufficiently small, namely $p \leq 0.188$. As mentioned in Section 2.2.1, this condition indicates that the difference $\mathbf{x}$ between two data points $\mathbf{h}$ and $\mathbf{h}'$ should contain a sufficient number of nonzero entries, suggesting that the two data points $\mathbf{h}$ and $\mathbf{h}'$ should be sufficiently distinct from each other. Then we can say that given the data distribution that exhibits sufficiently high discrimination between samples, the best classification performance can be attained using very sparse random matrices with sparsity $k = 1$, in terms of the maximum $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ achieved at $k = 1$.
- P2 implies that the classification performance will become comparable, as the matrix sparsity k increases. This is because as shown in equation 6, $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ will converge to a constant that merely depends on the data distribution and matrix size, as k tends to infinity. Furthermore, the convergence can be achieved when k is small. This is analyzed below. Note that in equation 6 we describe the convergence with $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ instead of $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$, 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 . The convergence error, namely the difference between the values of $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ with finite k and infinite k , is upper-bounded in equation 7, and the bound indicates a convergence speed $\mathcal{O}(1/\sqrt{k})$. By the bound equation 7, it is easy to further derive that

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

where η can be an arbitrary positive constant. It is seen that η sets an upper bound for the ratio between the convergence error with the convergence value (hereinafter referred to as the convergence ratio error). For any arbitrarily small upper bound η , as shown in equation 8, there exists a corresponding minimum sparsity k required to maintain this bound. When k falls within this specified range, $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ assumes similar values, suggesting that these k values should result in similar classification performance. Consequently, sparse matrices with small k values (around the minimum threshold) exhibit classification performance that is on par with denser matrices possessing larger k values. Note that the lower bound of k theoretically derived in equation 8 contains slack, and its actual value should be relatively small, typically on the order of tens, as evidenced by our numerical computations presented in Appendix C.1.

To verify the accuracy of our theoretical results P1 and P2, in Appendix C.1 we further investigate the trend of $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ in two ways: 1) by directly computing it with equation 9, and 2) by statistically estimating its value using synthetic data. Both approaches yield results consistent with P1 and P2. Statistical simulations also show that $\text{MAD}(\mathbf{z})$ and $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ share similar trends when varying matrix sparsity k , validating the equivalence relation between them as demonstrated in Section 2.3.

Finally, it is noteworthy that the theoretical results discussed above are derived based on the *expected* distance $\mathbb{E}\|\mathbf{R}\mathbf{x}\|_1$ (or equivalently, $m\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$), rather than the actual distance $\|\mathbf{R}\mathbf{x}\|_1$ that would be obtained with a single matrix. To approximate the expected distance and consequently attain the theoretical performance, a single matrix should have the size $m \geq \mathcal{O}(\sqrt{n})$ as shown in Property 2.

Property 2. Let $\mathbf{r}_i$ be the i -th row of a k -sparse random matrix $\mathbf{R} \in \{0, \pm\sqrt{\frac{n}{mk}}\}^{m \times n}$, and $\mathbf{x} \in \mathbb{R}^n$ with i.i.d. entries $x_i \sim \mathcal{T}(\mu, p, q)$. Suppose $z = \frac{1}{m}\|\mathbf{R}\mathbf{x}\|_1 = \frac{1}{m}\sum_{i=1}^m |\mathbf{r}_i^\top \mathbf{x}|$. For arbitrarily 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 $\mathbf{x}$.

3.2 GAUSSIAN MIXTURE DATA

Similarly to the analysis in the previous section, by examining $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ against varying matrix sparsity k , we investigate the relationship between $\text{MAD}(z)$ and k . Here, the original data vectors $\mathbf{h}$ and $\mathbf{h}'$ are drawn from Gaussian mixture distributions, such that their difference $\mathbf{x} = \mathbf{h} - \mathbf{h}'$ has i.i.d. entries $x_i \sim \mathcal{M}(\mu, \sigma^2, p, q)$ as specified in equation 2.

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

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

$$\begin{aligned} T_1 &= \sum_{i=0}^k C_k^i p^i q^{k-i} \left\lfloor \frac{k-i}{2} \right\rfloor C_{k-i}^{\lceil \frac{k-i}{2} \rceil} \\ T_2 &= \sum_{i=0}^k C_k^i p^i q^{k-i} \sum_{j=0}^{k-i} C_{k-i}^j e^{-\frac{(k-i-2j)^2 \mu^2}{2k\sigma^2}} \\ T_3 &= \sum_{i=0}^k C_k^i p^i q^{k-i} \sum_{j=0}^{k-i} C_{k-i}^j \Phi\left(-\frac{|k-i-2j|\mu}{\sqrt{k}\sigma}\right) \end{aligned}$$

and

$$\text{Var}(|\mathbf{r}^\top \mathbf{x}|) = \frac{n}{m}(\sigma^2 + 2q\mu^2) - (\mathbb{E}|\mathbf{r}^\top \mathbf{x}|)^2 \quad (10)$$

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

$$\mathbb{E}|\mathbf{r}^\top \mathbf{x}| \leq \mu\sqrt{\frac{n}{m}} + \sigma\sqrt{\frac{2n}{\pi m}}, \quad (11)$$

and

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

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

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

Remarks of Theorem 2: From the results, it can be seen that $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ exhibit similar trends as derived in P1 and P2 of Theorem 1. Specifically:

- Similarly to P1, by numerically computing equation 9 as elaborated in Appendix C.2, it can be observed that $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ will reach its maximum value at $k = 1$, when the data distribution parameter p (specified in equation 2) assumes relatively small values. This suggests that sparse random matrices can achieve the best classification performance at the sparsity level of $k = 1$, when the two data points sampled from the Gaussian-mixture distribution exhibit sufficient discriminability (corresponding to small p values).

- Similarly to P2, by equation 12 and equation 13, it can be deduced that $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ will converge to a constant as matrix sparsity k increases, with a convergence rate of $\mathcal{O}(1/\sqrt{k})$. With equation 13, the lower bound of k ensuring the convergence error ratio less than a given constant η can be derived:

$$\frac{\left| \frac{\sqrt{m}}{\mu\sqrt{n}} \mathbb{E}|\mathbf{r}^\top \mathbf{x}| - \sqrt{2(\sigma^2 + 2q\mu^2)/\pi} \right|}{\sqrt{2(\sigma^2 + 2q\mu^2)/\pi}} \leq \eta, \quad (14)$$

if $k \geq \left(\frac{4\sigma^3[p+2q(1+\mu^2/\sigma^2)^{3/2}]}{(\sigma^2+2q\mu^2)^{3/2}\sqrt{2}\eta} + \frac{3\sigma^4+2q(6\sigma^2\mu^2+\mu^4)}{(\sigma^2+2q\mu^2)\eta} \right)^2$. As remarked in Theorem 1, the lower bound of k derived in equation 14 for a given small η suggests a small matrix sparsity k that performs comparably to larger sparsity values in classification.

The resemblance between the results of Theorems 1 and 2 is not unexpected, as the ternary discrete distribution $x_i \sim \mathcal{T}(\mu, p, q)$ can be regarded as a limiting case of the three-component Gaussian mixture $x_i \sim \mathcal{M}(\mu, \sigma^2, p, q)$, where $\sigma \rightarrow 0$. Due to the excellent generalizability of Gaussian mixture models, the two properties discussed above apply to a diverse range of real-world data, as demonstrated in our experiments.

In Appendix C.2, we validate the results of Theorem 2 through numerical computations and statistical simulations. Similarly to the analysis of Theorem 1, we require the matrix size $m \geq \mathcal{O}(\sqrt{n})$, such that the actual distance $\|\mathbf{R}^\top \mathbf{x}\|_1$ for a specific matrix can approximate the expected distance $\mathbb{E}\|\mathbf{R}^\top \mathbf{x}\|_1$ (or equivalently, $m\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$) derived using equation 9. This ensures that the results of Theorem 2 which are based on $\mathbb{E}\|\mathbf{R}^\top \mathbf{x}\|_1$, also apply to $\|\mathbf{R}^\top \mathbf{x}\|_1$. The analysis is similar to that in Property 2, omitted here. By statistical simulations, we also see the similar trends of $\text{MAD}(\mathbf{z})$ and $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ with respect to variations in k , validating the equivalence relation between them as demonstrated in Section 2.3.

4 EXPERIMENTS

In this section, we aim to verify that impact of sparse matrices on classification is consistent with our theoretical predictions outlined in Theorems 1 and 2. More precisely, sparse matrices with only one or a few number of nonzero entries per row, can achieve classification performance comparable or even superior to that of denser matrices, particularly under the matrix size $m \geq \mathcal{O}(\sqrt{n})$.

4.1 SETTING

For the sake of generality, we evaluate our classification performance across datasets with diverse attributes and scales, including the YaleB image dataset (Georghiades et al., 2001; Lee et al., 2005), the Newsgroups text dataset (Joachims, 1997), the AMLALL gene dataset (Golub et al., 1999), the MNIST binary image dataset (Deng, 2012), the CIFAR100 image dataset (Krizhevsky & Hinton, 2009), as well as the large-scale ImageNet1000 image dataset (Krizhevsky et al., 2012). While most datasets can be approximately modeled using Gaussian mixtures, the MNIST dataset follows two-point distributions. For data details, refer to Appendix D.1.

To clearly reflect the separability of projected data, we adopt the naive K -nearest neighbor (KNN) classifier (Cover & Hart, 1967), which relies solely on pairwise data similarities without additional discrimination enhancement. For comprehensive validation of our theoretical results, we design two experimental settings: 1) binary classification on classical datasets (Figure 1), to verify basic performance; and 2) multiclass classification on large-scale datasets (Figure 2), specifically CIFAR100 and ImageNet1000, to test scalability under complex scenarios. Similarly to KNN , the desired performance trends can also be obtained with other more sophisticated classifiers, like SVMs (Cortes & Vapnik, 1995), as detailed in Appendix D.2. We also include Gaussian random projection as a baseline, given its prevalent use in the random projection research.

4.2 RESULTS

As shown in Figures 1 and 2, we evaluate the classification performance of sparse matrices across sparsity levels $k \in [1, 30]$ and projection ratios $m/n \in \{1\%, 10\%, 50\%\}$. The data dimension n is

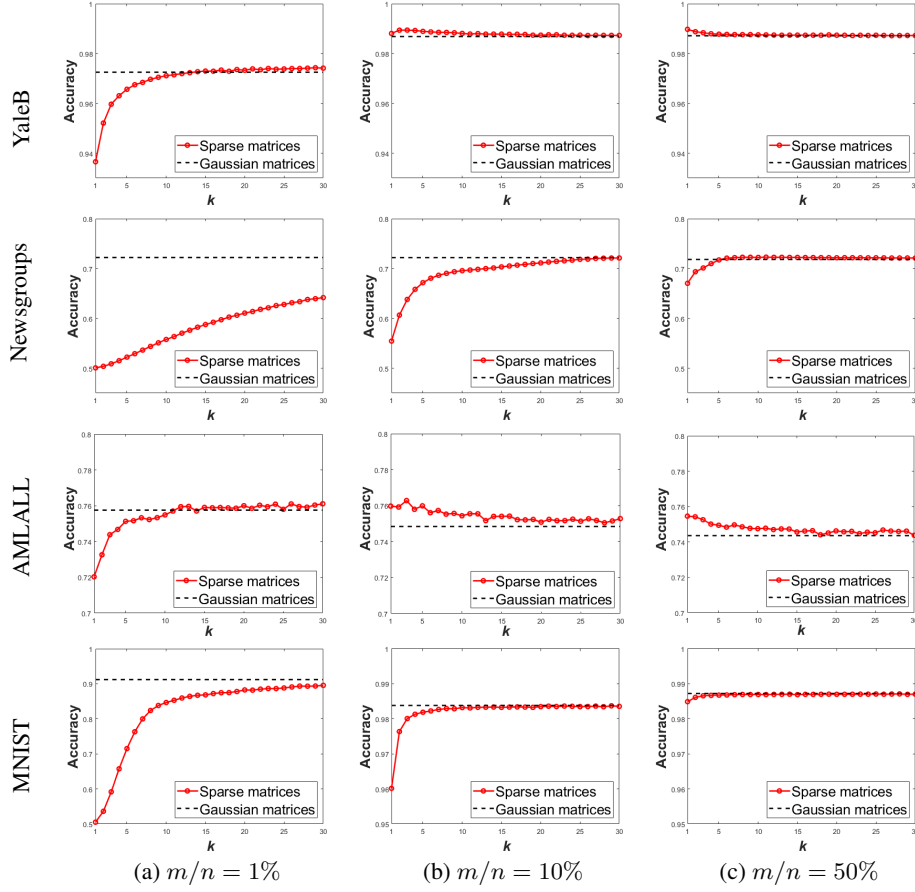


Figure 1: Classification accuracy of sparse matrix-based random projections with varying matrix sparsity $k \in [1, 30]$ and three different projection ratios $m/n = 1\%$, 10% and 50% . The data include the image data (YaleB, DCT features), text data (Newsgroups), gene expression data (AMLALL) and binary data (MNIST, binarized pixels). The performance of Gaussian matrix-based random projections is provided as a baseline.

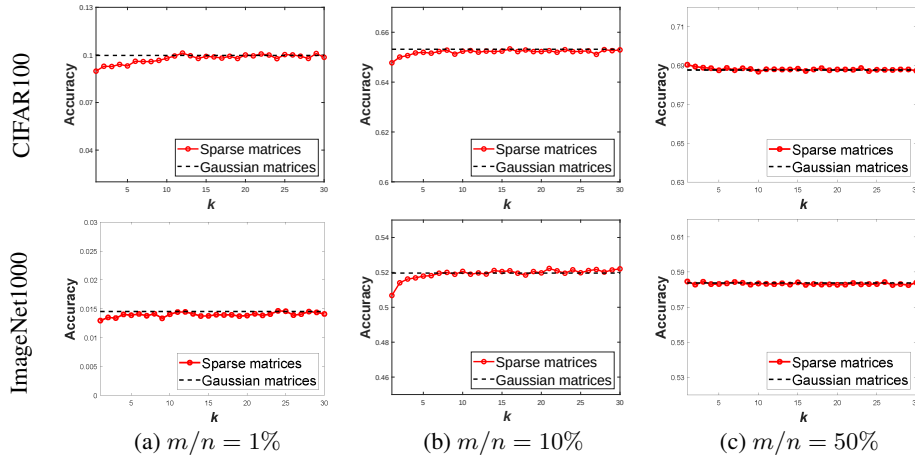


Figure 2: Multiclass classification accuracy of sparse matrix-based random projections with varying matrix sparsity $k \in [1, 30]$ and projection ratio $m/n \in \{1\%, 10\%, 50\%\}$. The datasets include CIFAR100 and ImageNet1000, with features extracted via ViT-B/32 (Dosovitskiy et al., 2020). The performance of Gaussian matrix-based random projections is provided as a baseline.

on the order of thousands. At this scale, the condition $m \geq \mathcal{O}(\sqrt{n})$ is naturally met for $m/n = 10\%$ and 50% , but not for $m/n = 1\%$.

These results reveal two main trends. First, classification accuracy stabilizes quickly as the projection ratio m/n increases. Convergence is achieved with relatively small k values ($k < 30$), when $m/n \geq 10\%$. Second, when $m/n \geq 50\%$, the best or near-best classification performance is typically attained at $k = 1$, i.e., when each row contains only one nonzero entry. Overall, the empirical trends in Figures 1 and 2 align closely with the theoretical patterns shown in Figures 3 and 6 (a,b) of Appendix C. These findings strongly support our theoretical conclusion: sparse matrices with only one or very few (i.e. $k < 30$) nonzero entries per row can match or exceed the performance of denser alternatives.

Furthermore, our experiments cover a wide range of datasets, from small ones like YaleB to large-scale ImageNet1000, and include diverse data types such as image features, text features, gene data, and binary-quantized data. This breadth thoroughly validates the universality and robustness of our theoretical insights.

Finally, sparse matrices demonstrate comparable and sometimes superior performance to Gaussian random matrices (the standard baseline), while offering significantly lower computational complexity. This makes them a compelling alternative to Gaussian matrices in real-world applications.

5 CONCLUSION

In our analysis of sparse $\{0, \pm 1\}$ -matrix-based random projections, we demonstrate that matrices with only one or few (such as less than thirty) nonzero entries per row can achieve comparable or even superior classification performance to denser alternatives. This theoretical result is consistent with classification experiments conducted across various datasets, ranging from small datasets like YaleB to the large dataset ImageNet1000, and encompassing a variety of data types, including image features, text features, gene data, and binary-quantized data. This showcases the broad applicability of our theoretical findings.

Moreover, our experimental results indicate that our extremely sparse matrices perform comparably to, and sometimes even better than, the commonly-used Gaussian matrices. Given the fundamental role and wide applications of random projection in machine learning for dimensionality reduction, the use of our sparse matrices can significantly reduce the computational complexity of related models, such as large-scale retrieval systems Charikar (2002), without compromising accuracy.

Furthermore, our study provides insights into the sparse structures inherent in more advanced models, such as deep networks (Li et al., 2016; Zhu et al., 2017; Wan et al., 2018; Marban et al., 2020; Rokh et al., 2023), where each layer can be modeled using random projections (Giryes et al., 2016).

ACKNOWLEDGMENTS

The language of the paper has been polished by large language models.

REFERENCES

- D. Achlioptas. Database-friendly random projections: Johnson–Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.
- 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.
- Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thirty-fourth annual ACM symposium on Theory of computing*, pp. 380–388, 2002.
- 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.
- Xinwei Deng, Ming Yuan, and Agus Sudjianto. A note on robust kernel principal component analysis. Contemporary Mathematics, 443:21–34, 2007.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.
- 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 Rameev 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.
- Mia Hubert, Tom Reynkens, Eric Schmitt, and Tim Verdonck. Sparse pca for high-dimensional data with outliers. Technometrics, 58(4):424–434, 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.
- A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Master’s thesis, Department of Computer Science, University of Toronto, 2009.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems, 2012.
- 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.
- Michael McCoy and Joel A Tropp. Two proposals for robust pca using semidefinite programming. Electronic Journal of Statistics, 5:1123–1160, 2011.
- Matthew P Mulè, Andrew J Martins, and John S Tsang. Normalizing and denoising protein expression data from droplet-based single cell profiling. Nature communications, 13(1):2099, 2022.
- 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.
- Simon Rowe and Andrew Blake. Statistical background modelling for tracking with a virtual camera. In BMVC, volume 95, pp. 423–432, 1995.
- 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.
- 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.
- Lizhe Wang, Hui Zhong, Rajiv Ranjan, Albert Zomaya, and Peng Liu. Estimating the statistical characteristics of remote sensing big data in the wavelet transform domain. IEEE Transactions on Emerging Topics in Computing, 2(3):324–337, 2014.
- 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.
- Eric P Xing, Michael I Jordan, and Richard M Karp. Feature selection for high-dimensional genomic microarray data. In Proceedings of the Eighteenth International Conference on Machine Learning, pp. 601–608, 2001.

- Ronald R. Yager and Naif Alajlan. A note on mean absolute deviation. Information Sciences, 279: 632–641, 2014. ISSN 0020-0255.
- 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.
- AS Zapevalov and AS Knyazkov. Parameterization of a two-component gaussian mixture for description of the sea surface. In Conference on Physical and Mathematical Modeling of Earth and Environment Processes, pp. 157–164. Springer, 2023.
- Chenzhuo Zhu, Song Han, Huizi Mao, and William J. Dally. Trained ternary quantization. In International Conference on Learning Representations, 2017.

A PROOFS FOR PROPERTY 1 IN SECTION 2

Proof. For a m -dimensional random vector $\mathbf{z}$ with i.i.d. entries $z_i \sim \mathcal{N}(\mu, \sigma^2)$, we have

$$\mathbb{E}\|\mathbf{z} - \mathbb{E}\mathbf{z}\|_1 = \mathbb{E}\left(\sum_{i=1}^m |z_i - \mathbb{E}z_i|\right) = \sum_{i=1}^m \mathbb{E}|z_i - \mathbb{E}z_i| = m\sqrt{\frac{2}{\pi}}\sigma \quad (15)$$

and

$$\mathbb{E}\|\mathbf{z} - \mathbf{z}'\|_1 = \mathbb{E}\left(\sum_{i=1}^m |z_i - z'_i|\right) = \sum_{i=1}^m \mathbb{E}|z_i - z'_i| = m\frac{2}{\sqrt{\pi}}\sigma \quad (16)$$

where $\mathbf{z}'$ is an independent copy of $\mathbf{z}$. Combining the above two results leads to

$$\mathbb{E}\|\mathbf{z} - \mathbb{E}\mathbf{z}\|_1 = \frac{1}{\sqrt{2}}\mathbb{E}\|\mathbf{z} - \mathbf{z}'\|_1 \quad (17)$$

□

B PROOFS FOR THEOREMS 1-2 AND PROPERTY 2 IN SECTION 3

B.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 $\mathbf{r}$ and $\mathbf{x}$, we can write $\|\mathbf{r}^\top \mathbf{x}\|_1 = \sqrt{\frac{n}{mk}}\mu \left| \sum_{i=1}^k z_i \right|$, where $z_i \in \{-1, 0, 1\}$ with probabilities $\{q, p, q\}$. Then, it can be derived that

$$\mathbb{E}|\mathbf{r}^\top \mathbf{x}| = \mu\sqrt{\frac{n}{mk}} \sum_{i=0}^k C_k^i p^i q^{k-i} \sum_{j=0}^{k-i} C_{k-i}^j |k-i-2j|, \quad (18)$$

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

$$\sum_{j=0}^{k-i} (C_{k-i}^j |k-i-2j|) = 2 \left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil}, \quad (19)$$

where $\lceil \alpha \rceil = \min\{\beta : \beta \geq \alpha, \beta \in \mathbb{Z}\}$. Combine (18) and (19), we can obtain equation 4.

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

$$\begin{aligned} \text{Var}(|\mathbf{r}^\top \mathbf{x}|) &= \text{Var}(\mathbf{r}^\top \mathbf{x}) - (\mathbb{E}|\mathbf{r}^\top \mathbf{x}|)^2 \\ &= \frac{2q\mu^2 n}{m} - \frac{4\mu^2 n}{mk} \left(\sum_{i=0}^k C_k^i p^i q^{k-i} \left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil} \right)^2. \end{aligned} \quad (20)$$

Proof of P1: This part aims to prove

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

where the subscript $k = 1$ denotes the case of $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ 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 $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ in terms of the two cases. For the case of $k = 1$, by equation 4, it is easy to derive that

$$\mathbb{E}|\mathbf{r}^\top \mathbf{x}|_{k=1} = 2q\mu\sqrt{\frac{n}{m}}. \quad (21)$$

Then, let us see the case of computing $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|_{k>1}$. By equation 4, $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|_{k>1}$ 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

$$\begin{aligned} & \frac{2}{\sqrt{k}} C_k^i p^i q^{k-i} \left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil} \\ & \leq \frac{1}{\sqrt{k}} C_k^i p^i q^{k-i} (k-i) 2^{k-i} \sqrt{\frac{2}{(k-i)\pi}} \\ & \leq \sqrt{\frac{2}{\pi}} C_k^i p^i (2q)^{k-i}, \end{aligned} \quad (22)$$

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

$$\begin{aligned} & \frac{2}{\sqrt{k}} C_k^i p^i q^{k-i} \left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil} \\ & \leq \frac{1}{\sqrt{k}} C_k^i p^i q^{k-i} (k-i) 2^{k-i} \sqrt{\frac{2}{(k-i-1)\pi}} \\ & = \sqrt{\frac{2}{\pi}} C_k^i p^i (2q)^{k-i} \frac{k-i}{\sqrt{k(k-i-1)}} \end{aligned} \quad (23)$$

Given $k \geq 5$, we further have

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

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,

$$\begin{aligned} & \frac{2}{\sqrt{k}} C_k^i p^i q^{k-i} \left\lceil \frac{k-i}{2} \right\rceil C_{k-i}^{\lceil \frac{k-i}{2} \rceil} \\ & \leq \begin{cases} \sqrt{\frac{2}{\pi}} C_k^i p^i (2q)^{k-i}, & k \geq 5, i \geq 2, \\ \frac{2}{\sqrt{k}} C_k^i p^i q^{k-i} (k-i) C_{k-i-1}^{\lceil \frac{k-i-1}{2} \rceil}, & \text{otherwise.} \end{cases} \end{aligned} \quad (24)$$

According to the results equation 22 and equation 24 derived in the above two cases, we know that $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|_{k>1}$ 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

$$\mathbb{E}|\mathbf{r}^\top \mathbf{x}| = \begin{cases} \frac{\mu\sqrt{n}}{\sqrt{2m}} (4q^2 + 4pq), & k = 2, \\ \frac{\mu\sqrt{n}}{\sqrt{3m}} (12q^3 + 12pq^2 + 6p^2q), & k = 3, \\ \frac{\mu\sqrt{n}}{\sqrt{m}} (12q^4 + 24pq^3 + 12p^2q^2 + 4p^3q), & k = 4, \end{cases} \quad (25)$$

and for the case of $k \geq 5$, with equation 22 and equation 24, we have

$$\mathbb{E}|\mathbf{r}^\top \mathbf{x}| \leq \mu \sqrt{\frac{2n}{\pi m}} + \mu \sqrt{\frac{n}{m}} (2q)^5 \left(\frac{3\sqrt{5}}{8} - \sqrt{\frac{2}{\pi}} \right). \quad (26)$$

By equation 21, equation 25 and equation 26, we can derive that

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

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

In what follows, we elaborate the proof of equation 26 by considering two cases of k , being even or odd.

Case 1: Suppose $k \geq 5$ and k is even. Combining equation 22 and equation 24, we have

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

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

$$h_1(k) = \frac{\sqrt{k}}{2^{k-1}} C_{k-1}^{\frac{k}{2}-1} \leq \lim_{k \rightarrow \infty} h_1(k) = \sqrt{\frac{2}{\pi}}. \quad (28)$$

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

$$\mathbb{E}|\mathbf{r}^\top \mathbf{x}| \leq \mu \sqrt{\frac{2n}{\pi m}}. \quad (29)$$

Case 2: Suppose $k \geq 5$ and k is odd. Combining (22) and (24), we have

$$\begin{aligned} \mathbb{E}|\mathbf{r}^\top \mathbf{x}| \leq & \mu \sqrt{\frac{n}{m}} C_k^0 (2q)^k \left(\frac{\sqrt{k}}{2^{k-1}} C_{k-1}^{\frac{k-1}{2}} - \sqrt{\frac{2}{\pi}} \right) \\ & + \mu \sqrt{\frac{2n}{\pi m}} \sum_{i=0}^k C_k^i p^i (2q)^{k-i}. \end{aligned} \quad (30)$$

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

$$h_2(k) = \frac{\sqrt{k}}{2^{k-1}} C_{k-1}^{\frac{k-1}{2}} \leq h_2(5) = \frac{\sqrt{5}}{2^4} C_4^2. \quad (31)$$

Then, it follows from (30) and (31) that

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

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

$$g(\mathbf{r}^\top \mathbf{x}; k, p) = \frac{\mathbb{E}|\mathbf{r}^\top \mathbf{x}|_k}{\mu\sqrt{n/m}} = \mathbb{E} \left| \frac{1}{\sqrt{k}} \sum_{i=1}^k z_i \right|, \quad (32)$$

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

$$\frac{1}{\sqrt{k}} \sum_{i=1}^k z_i \rightsquigarrow Z, \quad (33)$$

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

Then based on equation 26, 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

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

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

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

Next, let us investigate the error of the above convergence with respect to k . Following the definitions and properties described in equation 32 and equation 33, we further suppose $t_i = \frac{1}{\sqrt{2q}} z_i$ and $Q \sim N(0, 1)$, and get

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

where $d_w(\nu, v)$ denotes the Kolmogorov metric:

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

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

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

and then

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

□

B.2 PROOF OF PROPERTY 2

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

$$\begin{aligned} \mathbb{E}z^2 &= \mathbb{E}\left(\frac{1}{m} \sum_{i=1}^m |\mathbf{r}_i^\top \mathbf{x}|\right)^2 \\ &= \frac{1}{m^2} \mathbb{E}\left(\sum_{i=1}^m |\mathbf{r}_i^\top \mathbf{x}|^2\right) + \frac{1}{m^2} \mathbb{E}\left(\sum_{i \neq j} |\mathbf{r}_i^\top \mathbf{x}| \cdot |\mathbf{r}_j^\top \mathbf{x}|\right) \\ &= \frac{2q\mu^2 n}{m^2} + \frac{m-1}{2m} \mathbb{E}(|\mathbf{r}_i^\top \mathbf{x}| \cdot |\mathbf{r}_j^\top \mathbf{x}|). \end{aligned} \quad (34)$$

For the second term in the above result, it holds

$$\mathbb{E}(|\mathbf{r}_i^\top \mathbf{x}| \cdot |\mathbf{r}_j^\top \mathbf{x}|) \leq \text{Var}(|\mathbf{r}_i^\top \mathbf{x}|) + (\mathbb{E}|\mathbf{r}_i^\top \mathbf{x}|)^2 = \text{Var}(|\mathbf{r}_i^\top \mathbf{x}|) + (\mathbb{E}z)^2, \quad (35)$$

by the covariance property

$$\begin{aligned} \text{Cov}(|\mathbf{r}_i^\top \mathbf{x}|, |\mathbf{r}_j^\top \mathbf{x}|) &= \mathbb{E}(|\mathbf{r}_i^\top \mathbf{x}| \cdot |\mathbf{r}_j^\top \mathbf{x}|) - \mathbb{E}|\mathbf{r}_i^\top \mathbf{x}| \cdot \mathbb{E}|\mathbf{r}_j^\top \mathbf{x}| \\ &= \rho \sqrt{\text{Var}(|\mathbf{r}_i^\top \mathbf{x}|)} \cdot \sqrt{\text{Var}(|\mathbf{r}_j^\top \mathbf{x}|)} \\ &= \rho \text{Var}(|\mathbf{r}_i^\top \mathbf{x}|), \end{aligned} \quad (36)$$

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

Substituting equation 34 into $\text{Var}(z) = \mathbb{E}z^2 - (\mathbb{E}z)^2$, by the inequality equation 35 and equation 20, we can derive

$$\begin{aligned} \text{Var}(z) &\leq \frac{2q\mu^2 n}{m^2} + \frac{m-1}{2m} [\text{Var}(|\mathbf{r}_i^\top \mathbf{x}|) + (\mathbb{E}z)^2] - (\mathbb{E}z)^2 \\ &= \frac{2q\mu^2 n}{m^2} + \frac{m-1}{2m} \cdot \frac{2q\mu^2 n}{m^2} - (\mathbb{E}z)^2 \\ &= \frac{(m+1)q\mu^2 n}{m^2} - (\mathbb{E}z)^2. \end{aligned} \quad (37)$$

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

$$\Pr\{|z - \mathbb{E}z| \leq \varepsilon\} \geq 1 - \delta. \quad (38)$$

By the Chebyshev's Inequality, equation 38 will be achieved, if $\text{Var}(z)/\varepsilon^2 \leq \delta$; and according to equation 37, 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 $\mathbf{x}$. For a given $\mathbf{x}$, the condition of holding equation 38 can be further relaxed to $m^2 \geq \frac{2q\mu^2 n}{\varepsilon^2 \delta}$, since in this case $|\mathbf{r}_i^\top \mathbf{x}|$ is independent between different $i \in [m]$, such that $\text{Var}(z)$ changes to be equation 20 divided by m . □

B.3 PROOF OF THEOREM 2

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

$$\mathbb{E}|z| = \sqrt{\frac{2}{\pi}} \sigma e^{-\frac{\mu^2}{2\sigma^2}} + \mu \left(1 - 2\Phi\left(-\frac{\mu}{\sigma}\right)\right) \quad (39)$$

which will be used in the sequel. With the distributions of $\mathbf{r}$ and $\mathbf{x}$, we have $|\mathbf{r}^\top \mathbf{x}| = \sqrt{\frac{n}{mk}} \left| \sum_{i=1}^k x_i \right|$. 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 equation 39 we can derive that

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

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

The above equation and equation 21, equation 25, equation 26 together lead to

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

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

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

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

C NUMERICAL VALIDATIONS OF THEOREMS 1 AND 2

C.1 NUMERICAL VALIDATION OF THEOREM 1

C.1.1 VALIDATING P1 AND P2 BY DIRECTLY COMPUTING $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ WITH EQUATION 4

To more accurately examine the changing trend of $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ against varying matrix sparsity k (derived in P1 and P2), we directly compute the value of $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ by equation 4. Note that besides the parameter k , $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ also involves the parameter p , the probability of $x_i = 0$ as specified in equation 1. So we investigate $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ 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 Figures 3 (a) and (b). The results exhibit two properties similar to those predicted by P1 and P2:

- (P3) When $p \leq 1/3$, such as the case of $p = 1/3$ shown in Figure 3(a), $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ 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 Figure 3(b). The results indicate that to maximize $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ at $k = 1$, the condition $p \in [0, 1/3]$ is sufficient, which is broader than the theoretical requirement $p \in [0, 0.188]$ derived from P1. Recall that a wider range of p allows for a larger space of data as modeled in (3). This suggests that the desired property of maximizing $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ at $k = 1$ can be achieved over a wider range of p values than what was theoretically predicted. To achieve a small p within the range of $p \in [0, 1/3]$, as pointed out in Section 2.2.2, the original data points $\mathbf{h}$ and $\mathbf{h}'$ need to exhibit sufficiently high discrimination between them.
- (P4) With the increasing of k , as the two cases of $p = 1/3$ and $2/3$ shown in Figures 3(a) and (b), $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ tends to converge to the limit value $2\sqrt{q/\pi}$ derived in equation 6, where $q = (1 - p)/2$. Furthermore, it can be seen that small convergence errors will be achieved, when k is very small, typically in the range of a few tens. For instance, in Figure 4(a) we derive the convergence error ratios as defined in equation 8, which give the values close to zero when $k \geq 20$ and p is relatively small. Recall that the small p value implies that the original data have high discrimination between each other. With the decreasing of data discrimination, we should need larger k to achieve small convergence errors.

In the analysis of the expected distance $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$, the influence of the variance $\text{Var}(|\mathbf{r}^\top \mathbf{x}|)$ in equation 5 should be considered. Statistically, a lower variance $\text{Var}(|\mathbf{r}^\top \mathbf{x}|)$ indicates a higher probability that the actual distance $|\mathbf{r}^\top \mathbf{x}|$ of a single matrix closely approximates its expected value $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$. Also, this implies a higher consistence between theoretical and practical results. By computing equation 5, we observe a trend similar to $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$: as k increases, $\text{Var}(|\mathbf{r}^\top \mathbf{x}|)$ tends to quickly converge to a constant value. This suggests that $\text{Var}(|\mathbf{r}^\top \mathbf{x}|)$ varies minimally across different k values. Therefore, the probability of $|\mathbf{r}^\top \mathbf{x}|$ approximating $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ remains consistent for various k , enabling us to use $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ to reasonably estimate and compare the distances $|\mathbf{r}^\top \mathbf{x}|$ of actual matrices across different k .

C.1.2 VALIDATING P1 AND P2 BY STATISTICALLY ESTIMATING $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ WITH SYNTHETIC DATA

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

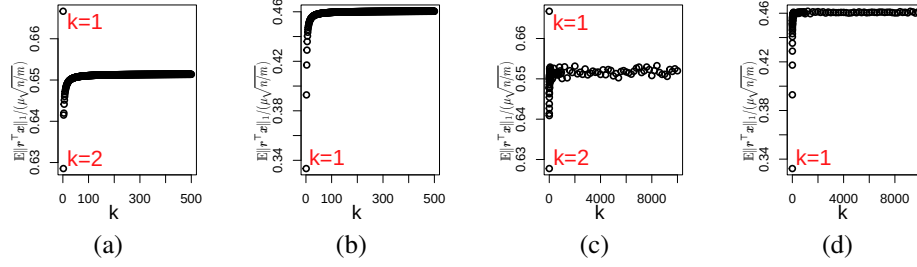


Figure 3: The value of $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ calculated by equation 4 with $p = 1/3$ (a) and $p = 2/3$ (b), and estimated by statistical simulation with $p = 1/3$ (c) and $p = 2/3$ (d), provided $x_i \sim \mathcal{T}(\mu, p, q)$, $\mu = 1$.

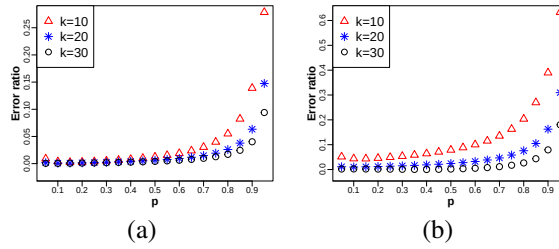


Figure 4: The convergence error ratios of three different $k \in \{10, 20, 30\}$ over varying p are derived for two-point distributed data (a) and Gaussian mixture data (b), by computing the left side of the inequality of η shown respectively in equation 8 and equation 14.

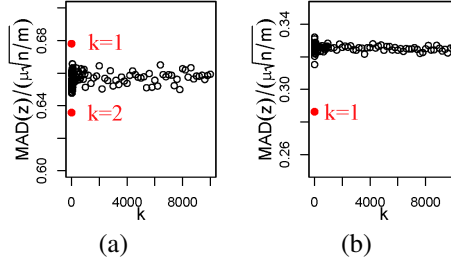


Figure 5: The value of $\text{MAD}(\mathbf{z})$ against varying k is estimated in (a) and (b), respectively, with the synthetic data generated using the parameters $p = 1/3$ and $p = 2/3$ as defined in Figures 3 (c) and (d) for $x_i \sim \mathcal{T}(\mu, p, q)$, $\mu = 1$.

that the choices of m , n and μ will not affect the changing trend of $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ against k . Comparing the numerical computation results with the simulation results presented in Figure 3, specifically contrasting (a) vs. (c) and (b) vs. (d), it is seen that both sets of results exhibit similar trends in the variation of $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$. The similarity between them validates Theorem 1, as well as the numerical computation results P3 and P4.

C.1.3 VALIDATING EQUIVALENCE RELATIONS BETWEEN $\text{MAD}(\mathbf{z})$ AND $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$

In Section 2.3, our statistical analysis demonstrates that the trend of $\text{MAD}(\mathbf{z})$ against varying matrix sparsity k can be estimated with $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$. By comparing the simulation values of $\text{MAD}(\mathbf{z})$ (Figures 5 (a) and (b)) with those of $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ (Figures 3 (c) and (d)), it is seen that they indeed exhibit similar trends. This validates the equivalence relation between them.

C.2 NUMERICAL VALIDATION OF THEOREM 2

C.2.1 VALIDATING EQUATION 12 BY DIRECTLY COMPUTING $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ WITH EQUATION 9

In this part, we directly compute the value of $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ by equation 9. Note that $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ involves four parameters: k , p , μ , and σ . In computing equation 9, we fix $\mu = 1$ and vary other parameters in the ranges of $\sigma/\mu \in (0, 1/3)$, $p \in (0, 1)$ and $k \in [1, 500]$. Here, we upper bound the value range of σ/μ by $1/3$ for easy simulation. Empirically, the changing trend of $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ is not sensitive to the varying of σ/μ , but sensitive to p , i.e. the probability of each entry x_i of the data difference $\mathbf{x}$ taking zero value, as specified in equation 2. In Figures 6(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 Figure 6(a), $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ 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 Figure 6(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 enable obtaining the maximum $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ at $k = 1$.
- (P6) With the increasing of k , as the two results shown in Figure 6(a) and (b), $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ converges to the limit value derived in equation 12. Similarly to the convergence discussed in P4 for two-point distributed data, the convergence error ratio defined in equation 14 can approach zero with small k , such as $k = 20$ shown in Figure 4(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 ternary discrete distribution $x_i \sim \mathcal{T}(\mu, p, q)$ can be viewed as an extreme case of the three-component Gaussian mixture $x_i \sim \mathcal{M}(\mu, \sigma^2, p, q)$ with $\sigma \rightarrow 0$. Thanks to the good generalizability of Gaussian mixture models, as will be seen in our experiments, the two properties analyzed above apply to a variety of real-world data.

C.2.2 VALIDATING EQUATION 12 BY STATISTICALLY ESTIMATING $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/(\mu\sqrt{n/m})$ WITH SYNTHETIC DATA

Similarly as in Section C.1.2, we here verify the accuracy of Theorem 2, including the expression equation 9 of $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ and its convergence equation 12 by performing statistical simulations on $\mathbf{x}$ and $\mathbf{r}$. The simulation results should agree with the numerical computation results P5 and P6, if the theorem is correct. In the simulation, we estimate the value of $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|/\sqrt{n/m}$ by drawing 10^6 pairs of $\mathbf{x}$ and $\mathbf{r}$ from their respective distributions and then computing the average of $|\mathbf{r}^\top \mathbf{x}|_1/\sqrt{n/m}$ as the estimate. The parameters of the distributions of $\mathbf{x}$ and $\mathbf{r}$ 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 k to vary between 1 and 10000. The average value of $|\mathbf{r}^\top \mathbf{x}|/\sqrt{n/m}$ at each k is presented in Figures 6(c) and (d), with $p = 1/2$ and $2/3$, respectively. Comparing the numerical computation results and the simulation results shown in Figure 6, specifically 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 computation results P5 and P6.

C.2.3 VALIDATING EQUIVALENCE RELATIONS BETWEEN $\text{MAD}(z)$ AND $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$

In Section 2.3, our statistical analysis demonstrates that the trend of $\text{MAD}(z)$ against varying matrix sparsity k can be estimated with $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$. By comparing the simulation values of $\text{MAD}(z)$ (Figures 7 (a) and (b)) with those of $\mathbb{E}|\mathbf{r}^\top \mathbf{x}|$ (Figures 6 (c) and (d)), it is seen that they indeed present similar trends. This validates the equivalence relation between them.

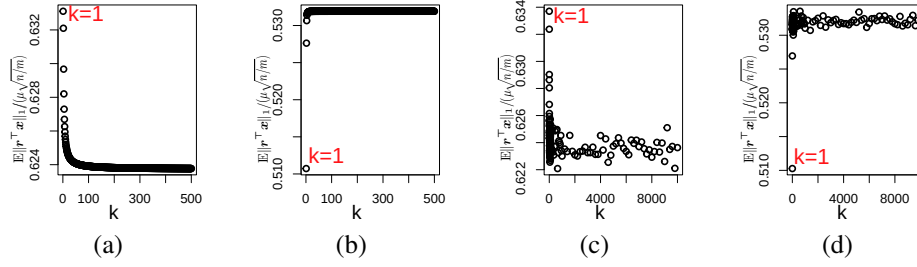


Figure 6: The value of $\mathbb{E}||r^T x||_2 / \sqrt{n/m}$ calculated by equation 9 with $p = 1/2$ (a) and $p = 2/3$ (b), and estimated by statistical simulation with $p = 1/2$ (c) and $p = 2/3$ (d), provided $x_i \sim \mathcal{M}(p, q, \mu, \sigma^2)$, $\mu = 1$ and $\sigma = 1/3$.

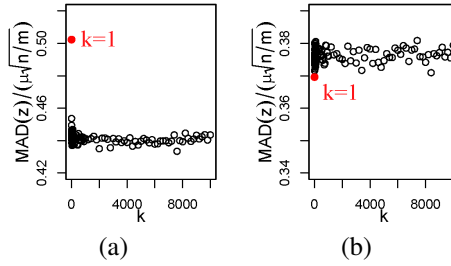


Figure 7: The value of $\text{MAD}(z)$ against varying k is estimated in (a) and (b), respectively, with the synthetic data generated using the parameters $p = 1/2$ and $p = 2/3$ as defined in Figures 6 (c) and (d) for $x_i \sim \mathcal{M}(p, q, \mu, \sigma^2)$, $\mu = 1$ and $\sigma = 1/3$.

D EXPERIMENTS

D.1 SETTING

D.1.1 DATA

For the sake of generality, we evaluate our classification performance across datasets with diverse attributes and scales, including the YaleB image dataset (Georgiades et al., 2001; Lee et al., 2005), the Newsgroups text dataset (Joachims, 1997), the AMLALL gene dataset (Golub et al., 1999), the MNIST binary image dataset (Deng, 2012), the CIFAR100 image dataset (Krizhevsky & Hinton, 2009), as well as the large-scale ImageNet1000 image dataset (Krizhevsky et al., 2012). While most datasets can be approximately modeled using Gaussian mixtures, the MNIST dataset follows two-point distributions. The data settings are introduced as follows. YaleB contains 168×192 -sized face images of 38 persons, with about 64 faces per person. For easier simulation, we reduce the image size to 40×30 . Newsgroups consists of 20 categories of text data, with 500 samples per category. Each sample is represented with a 3060-dimensional bag-of-words feature vector. 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. CIFAR100 contains 100 categories of 32×32 -sized color images, with 600 samples for each category. Among them, 5/6 of samples are used as the training set. ImageNet1000 contains 1,000 object categories, with approximately 1,000 images for each category. In total, there are about 1.2 million training images, 50,000 validation images, and 100,000 unlabeled test images. For ImageNet1000 and CIFAR100, we extract their features using the Vision-Transformer model ViT-B/32 (Dosovitskiy et al., 2020).

D.1.2 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 clearly reflect the separability of projected data, we adopt the naive K -nearest neighbor (KNN) classifier (with $K = 5$) (Cover & Hart, 1967), which relies solely on pairwise data similarities without additional discrimination enhancement. For comprehensive validation of our theoretical results, we design two experimental settings: 1) binary classification on classical datasets (Figure 1), to verify basic performance; and 2) multiclass classification on large-scale datasets (Figure 2), specifically CIFAR100 and ImageNet1000, to test scalability under complex scenarios. Similarly to KNN , the desired performance trends can also be obtained with other more sophisticated classifiers, like SVMs (Cortes & Vapnik, 1995), as detailed below. We also include Gaussian random projection as a baseline, given its prevalent use in the random projection research.

In the binary classification, we enumerate all possible class pairs in each dataset. For each class of data, 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 voting. Usually, each classification process takes less than 0.0001 seconds on a computer equipped with an Intel Core i9-10980XE CPU and 256G of RAM. The multiclass classification adopts the testing and training sets defaulted in CIFAR100 and ImageNet1000.

D.2 SVM CLASSIFICATION RESULTS

In Figure 8, we test the SVM (with linear kernel) classification accuracy for sparse matrices with varying matrix sparsity k and projection ratio m/n on four different types of datasets. It is 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.

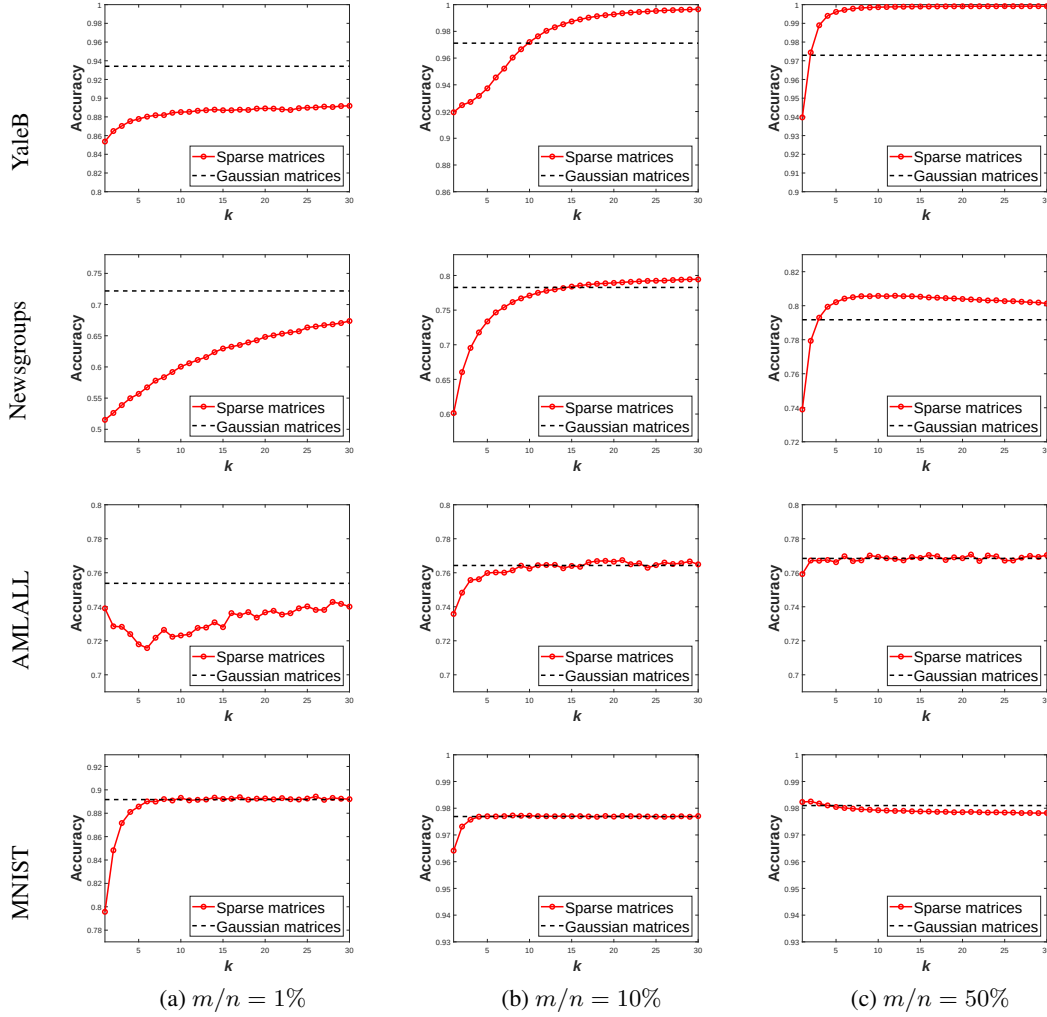


Figure 8: SVM classification accuracy of sparse matrix-based random projections with varying matrix sparsity $k \in [1, 30]$ and three different projection ratios $m/n = 1\%$, 10% and 50% . The data include the image data (YaleB, DCT features), text data (Newsgroups), gene expression data (AMLALL) and binary data (MNIST, binarized pixels). The performance of Gaussian matrix-based random projections is provided as a baseline.