Minimax Gaussian Classification & Clustering

Tianyang Li[1] **Xinyang Yi**[1] **Constantine Caramanis**[1] **Pradeep Ravikumar**[2]
[1]University of Texas, Austin [2]Carnegie Mellon University

Abstract

We present minimax bounds for classification and clustering error in the setting where covariates are drawn from a mixture of two isotropic Gaussian distributions. Here, we define clustering error in a *discriminative* fashion, demonstrating fundamental connections between classification (supervised) and clustering (unsupervised). For both classification and clustering, our lower bounds show that without enough samples, the best any classifier or clustering rule can do is close to random guessing. For classification, as part of our upper bound analysis, we show that Fisher's linear discriminant achieves a fast minimax rate $\Theta(1/n)$ with enough samples n. For clustering, as part of our upper bound analysis, we show that a clustering rule constructed using principal component analysis achieves the minimax rate with enough samples. We also provide lower and upper bounds for the high-dimensional sparse setting where the dimensionality of the covariates p is potentially larger than the number of samples n, but where the difference between the Gaussian means is sparse.

1 Introduction

We consider lower bounds and upper bounds for Gaussian classification and clustering. We focus on the setting with two classes/clusters of isotropic (spherical) Gaussian distributions. Our results reveal the dependency of classification and clustering errors on the dimension of the covariates/features p, number of samples n, sparsity of the optimal classifier/clustering rule $s = \|\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\|_0$, and separation between two classes/clusters $\rho = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}$. Note that ρ is invariant under linear transforms of the features, and we are interested in the difficult case when $\rho = O(1)$ is small.

Gaussian classification [11] is the problem of labeling a new sample using its features, given already observed samples with both the label and the features, where we have a prior distribution over the labels, and each class's features are distributed according to a Gaussian distribution. The classification error of classifier is defined as the probability that the classifier mislabels a sample, which we also term as the risk of the classifier. When analyzing classification algorithms, we are interested in bounding the excess risk, which quantifies how much worse the trained classifier is compared to the optimal classifier – the Bayes classifier. In this paper we present a fast $\Theta(1/n)$ minimax rate for Gaussian classification, even without the classes being completely separable (the Bayes error is not 0). For lower bounds, we also show that without enough samples, the performance of any classifier is close to random guessing. Random guessing uniformly randomly assigns a sample into one of the two classes, and its classification error is 1/2. For upper bounds, we show that Fisher's linear discriminant [11] achieves the minimax rate for classification of two balanced isotropic Gaussian distributions. In the sparse setting, we use a ℓ_1 regularized optimization problem to estimate the Bayes classifier.

Gaussian clustering [11] is the problem of labeling a new sample using its features, given already observed *unlabeled* samples with only the features. The features are distributed according to a mixture of two Gaussian distributions. We define clustering error in a *discriminative* fashion, and in the spirit of classification error: we define it as the minimum classification error over the two mappings of the two clusters to the two labels. It is easy to see that the Bayes classifier is the optimal clustering rule, and the clustering error is equal to the Bayes classification error. We show that a clustering rule constructed using principal component analysis achieves a fast $\Theta(1/n)$ minimax rate. In the sparse setting, we first use a thresholded estimate of the optimal clustering rule's support, and then apply principal component analysis to estimate the optimal clustering rule.

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54.

We would like to note that when the separation between the two classes/clusters is constant, our results for classification and clustering are tight up to logarithmic factors in dimension p, number of samples n, and sparsity s. One of our key contributions is the derivation of classification and clustering minimax risk lower bounds. This presents particular technical challenges, since unlike metrics, risk lower bounds do not trivially satisfy a triangle inequality (which is critically used in typical lower bound derivations).

Related Work Until recently, results giving a $\Theta(1/n)$ minimax rate relied critically on a strict separation assumption (the Bayes classifier's classification error is 0) [15]. Without this assumption, the best known results were $O(1/\sqrt{n})$ classification error rates [21, 6], or $o(1/\sqrt{n})$ minimax rates[17, 14]. For certain nonparametric problems, matching upper and lower bounds [17] for classification excess risk are known. However, matching upper and lower bounds in parametric problems are not known, although Theorem 13.21 of [5] gives a $o(1/\sqrt{n})$ fast rate upper bound for parametric classification problems under a VC dimension condition and Tsybakov's low noise condition. In a recent result, [16] showed that it is possible to achieve a $O(1/n)$ rate for excess risk in Gaussian classification without such a separation assumption. The question of a lower bound for this setting remained open. Resolving this and proving a matching lower (minimax) bound is one of the contributions of this paper.

For clustering, most previous works have focused on recovering each cluster [19, 10]. Despite other discriminative clustering methods without theoretical guarantees [22, 13], only [1] has formally defined clustering error. Our definition of clustering error draws fundamental connections between classification (supervised) and clustering (unsupervised), and in particular, is different from that of [1]. Using the notations of Section 2.2, for a clustering rule C, [1] defines the clustering error as $\min\{\Pr[\mathsf{C} \neq \mathsf{C}^*], 1 - \Pr[\mathsf{C} \neq \mathsf{C}^*]\}$, where C^* is the clustering rule using the Bayes classifier. In this paper, we define the clustering error as $\min\{\Pr[\mathsf{C} \neq Y], 1 - \Pr[\mathsf{C} \neq Y]\}$, where Y is the latent indicator in the Gaussian mixture model. In the former definition, the risk of an optimal classification rule is always zero, and hence does not reveal the inherent "hardness" of the problem, which (as is intuitive) is characterized by the distance between the cluster centers – a quantity we denote by ρ in the sequel. Conversely, defining the clustering error as we do, we obtain results that explicitly depend on ρ. Section 4.3 gives a more detailed comparison.

2 Problem Setup

2.1 Classification

The Gaussian classification problem is specified as follows. Suppose we use $\boldsymbol{X} \in \mathbb{R}^p$ to denote the features, and use $Y \in \{0, 1\}$ to denote the label. Then joint distribution over the features and the label is then specified by a simple Bernoulli distribution for the prior distribution over Y, and that the features conditioned on each of the two labels are distributed according to Gaussian distributions, with the same covariance matrix but different means. We thus have:

$$\boldsymbol{X} \mid Y \sim \mathcal{N}(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}), \tag{1}$$

$$\Pr[Y = 1] = \pi_1,\ \Pr[Y = 0] = \pi_0. \tag{2}$$

A classifier is a function $\mathsf{C} : \mathbb{R}^p \to \{0, 1\}$. A linear classifier with parameter $\boldsymbol{w} \in \mathbb{R}^p$, $b \in \mathbb{R}$ is defined as $\mathbf{1}(\boldsymbol{w}^T \boldsymbol{x} + b > 0)$.

The classification error of a classifier C is defined as the probability that the classifier mislabels a sample

$$\mathcal{R}(\mathsf{C}) = \Pr[\mathsf{C}(\boldsymbol{X}) \neq Y]. \tag{3}$$

For a linear classifier $\mathsf{C}(\boldsymbol{x}) = \mathbf{1}(\boldsymbol{w}^T \boldsymbol{x} + b > 0)$, the classification error can be written as

$$1 - \pi_1 \Phi(\tfrac{\boldsymbol{w}^T \boldsymbol{\mu}_1 + b}{\sqrt{\boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w}}}) - \pi_0 \Phi(-\tfrac{\boldsymbol{w}^T \boldsymbol{\mu}_0 + b}{\sqrt{\boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w}}}). \tag{4}$$

The optimal classifier is the Bayes classifier $\mathsf{C}^*(\boldsymbol{x}) = \mathbf{1}(\frac{\pi_1 \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})}{\pi_0 \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})} > 1)$, which is a linear classifier

$$\begin{aligned} &\mathsf{C}^*(\boldsymbol{x}) = \mathbf{1}((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} \\ &m + \frac{1}{2}(-\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\mu}_0^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_0) + \log \frac{\pi_1}{\pi_0} > 0). \end{aligned} \tag{5}$$

A classifier is trained using n i.i.d. samples $(\boldsymbol{x}_1, y_i), \ldots, (\boldsymbol{x}_n, y_n)$, where we have access to both the features and the label.

2.2 Clustering

The Gaussian clustering problem is specified as follows. Suppose we use $\boldsymbol{X} \in \mathbb{R}^p$ to denote the features. No label is observed. The samples are distributed according to a Gaussian mixture model with two mixture components, each with the same covariance matrix but different means:

$$\boldsymbol{X} \sim \pi_1 \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + \pi_0 \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}). \tag{6}$$

A clustering rule is a function $\mathsf{C} : \mathbb{R}^p \to \{0, 1\}$. A linear clustering rule with parameter $\boldsymbol{w} \in \mathbb{R}^p$, $b \in \mathbb{R}$ is defined as $\mathbf{1}(\boldsymbol{w}^T \boldsymbol{x} + b > 0)$. Note that, in the clustering problem, we can view the classification label Y as a latent variable, and in particular, it is natural to use an error metric similar to classification error as clustering error. Here, we define the clustering error of a clustering rule C as

$$\mathcal{R}(\mathsf{C}) = \min\{\Pr[\mathsf{C}(\boldsymbol{X}) \neq Y],\ \Pr[\mathsf{C}(\boldsymbol{X}) \neq 1 - Y]\}, \tag{7}$$

where a minimum is taken over two possible labelings because there is no access to the true label.

For a linear clustering rule $\mathsf{C}(\boldsymbol{x}) = \mathbf{1}(\boldsymbol{w}^T\boldsymbol{x} + b > 0)$, the clustering error can be written as

$$\begin{aligned}\min\{&\pi_1\Phi(\tfrac{\boldsymbol{w}^T\boldsymbol{\mu}_1+b}{\sqrt{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}}}) + \pi_0\Phi(-\tfrac{\boldsymbol{w}^T\boldsymbol{\mu}_0+b}{\sqrt{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}}}),\\ &\pi_1\Phi(-\tfrac{\boldsymbol{w}^T\boldsymbol{\mu}_1+b}{\sqrt{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}}}) + \pi_0\Phi(\tfrac{\boldsymbol{w}^T\boldsymbol{\mu}_0+b}{\sqrt{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}}})\}.\end{aligned} \tag{8}$$

It is easy to see that using the Bayes classifier as a linear clustering rule minimizes the clustering error, and this error is equal to the Bayes classification error.

A clustering rule is trained using n i.i.d. samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Here we only have access to the features, but not the latent indicator (the label).

2.3 Excess risk

Note that in both classification and clustering, the Bayes classifier minimizes the error, and the minimum error is the Bayes classification error. Taking this into consideration, for a given Gaussian classification or clustering problem, we define the excess risk of a classifier or clustering rule C as

$$\mathcal{E}(\mathsf{C}) = \mathcal{R}(\mathsf{C}) - \mathcal{R}^*, \tag{9}$$

where $R^* = 1/2 - 1/2\int|\pi_1\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) - \pi_0\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})|\ d\boldsymbol{x} > 0$ is the Bayes classification error.

2.4 Assumptions

In this paper, we consider classification and clustering of two balanced isotropic Gaussian distributions. In classification, we consider distributions of the type $\Pr[Y = 1] = \Pr[Y = 0] = 1/2$ and $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}_Y, \sigma^2\mathbf{I})$. In clustering, we consider distributions of the type $\frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_1, \sigma^2\mathbf{I}) + \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_0, \sigma^2\mathbf{I})$. Our setting is similar to those presented in [4, 8, 1]. We will present results in the general setting without any sparsity, and then results with the following sparsity assumption $s = \|\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)\|_0 = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|_0$.

3 Bounds for Gaussian Classification

In this section, we present lower and upper bounds for Gaussian classification, both in the general setting without sparsity, and in the high-dimensional sparse setting where the difference between the means is sparse.

Our lower bounds show that when there are not enough samples, the best any classifier can do is close to random guessing. For the general setting, we show that Fisher's linear discriminant method achieves the minimax rate up to constant factors. Note that the constants in the bounds can be improved.

3.1 General Setting without Sparsity

3.1.1 Lower Bound

For the lower bound, we consider Gaussian classification problems where $\Pr[Y = 1] = \Pr[Y = 0] = 1/2$, $\boldsymbol{X} \mid Y \sim \mathcal{N}(\boldsymbol{\mu}_Y, \mathbf{I})$ with $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_0 = \boldsymbol{\mu}$. In this case, the Bayes classification error is $1 - \Phi(\|\boldsymbol{\mu}\|_2)$. We will index each classification problem with $\boldsymbol{\mu}$.

The following theorem provides a lower bound for Gaussian classification when enough samples are available.

Theorem 1 (Classification excess risk lower bound)**.** *Let $\rho > 0$ be a fixed number. For sufficiently large p and n, and any classifier C trained using n samples, we have*

$$\max_{\|\boldsymbol{\mu}\|_2=\rho/2} \mathbb{E}_{\boldsymbol{\mu}}[\mathcal{E}_{\boldsymbol{\mu}}(\mathsf{C})] \gtrsim e^{-\rho^2/8}\min\{\tfrac{1}{\rho}\tfrac{p}{n}, \rho\}. \tag{10}$$

The proof of Theorem 1 is based on the following theorem for general classification problems, which provides a "triangle inequality" in Fano's method [25] for a fixed classifier's excess risk in different classification problems.

Theorem 2 (A "triangle inequality" for classification excess risk)**.** *Let i correspond to a classification problem where $\Pr[Y = y] = \pi_y^{(i)}$ and $\boldsymbol{X} \mid Y \sim p_Y^{(i)}(\boldsymbol{X})$. Denote the Bayes classifier as $\mathsf{C}^{*(i)}(\boldsymbol{x}) = \mathbf{1}((\pi_1^{(i)}p_1^{(i)}(\boldsymbol{x}))/(\pi_0^{(i)}p_0^{(i)}(\boldsymbol{x})) > 1)$. Suppose we have another classification problem j. Then for any fixed classifier C,*

$$\begin{aligned}\mathcal{E}_i(\mathsf{C}) + \mathcal{E}_j(\mathsf{C}) \geq \int\limits_{\mathsf{C}^{*(i)}\neq\mathsf{C}^{*(j)}} &\min\{|\pi_1^{(i)}p_1^{(i)} - \pi_0^{(i)}p_0^{(i)}|,\\ &|\pi_1^{(j)}p_1^{(j)} - \pi_0^{(j)}p_0^{(j)}|\}\ d\boldsymbol{x}.\end{aligned} \tag{11}$$

Indeed, for a classification problem where $\Pr[Y = y] = \pi_y$ and $\boldsymbol{X} \mid Y \sim p_Y(\boldsymbol{X})$, a classifier C's excess risk $\mathcal{E}(\mathsf{C}) = \int_{\mathsf{C}\neq\mathsf{C}^*}|\pi_1p_1 - \pi_0p_0|\ d\boldsymbol{x}$ where C^* is the Bayes classifier (Theorem 2.2, [7]).

To construct a packing for applying Fano's method, we use the following lemma on existence of sparse sets (Lemma 4.10, [18]). This lemma is also known as the Gilbert-Varshamov bound (Theorem 17.2, [12]). It is stated here for completeness.

Lemma 1 (Existence of sparse sets (Lemma 4.10, [18]))**.** *Let* $\Psi = \{\psi \in \{-1,+1\}^p : \|\psi\|_0 = s\}$ *for positive integers* p *and* $1 \le s < p/4$*. Then there exists* $\psi_1, \dots, \psi_N$ *such that the Hamming distance* $\delta(\psi_i, \psi_j) > s/2$ *for all* $1 \le i < j \le N$*, and* $\log N \ge \frac{s}{5}\log\frac{p}{s}$*.*

We briefly sketch Theorem 1's proof. Let us consider the set $M = \{\boldsymbol{\mu} \mid (\sqrt{\rho^2/4 - s\alpha^2}), \pm\alpha\psi_1, \dots, \pm\alpha\psi_{p-1}) \in \mathbb{R}^p\}$ where $s = (p-1)/6 > p/8$, and $\psi_1, \dots, \psi_{p-1}$ are as given in Lemma 1. And we set $\alpha = 0.001\min\{\sqrt{1/n}, \rho/(2\sqrt{s})\}$. Using Theorem 2 we can show that for any fixed $\boldsymbol{\mu}$, if $\mathcal{E}_{\boldsymbol{\mu}} > 4 \times 10^{-9}\frac{e^{-\rho^2/8}}{\rho}s\alpha^2$ then we must have $\mathcal{E}_{\boldsymbol{\mu}'} < 4 \times 10^{-9}\frac{e^{-\rho^2/8}}{\rho}s\alpha^2$ for all other $\boldsymbol{\mu}' \neq \boldsymbol{\mu}$. This observation reduces the problem to a testing problem. So using Fano's method, we can show that

$$
\begin{aligned}
&\Pr[\mathcal{E}_{\boldsymbol{\mu}}(\mathsf{C}) \ge 4 \times 10^{-9}\frac{e^{-\rho^2/8}}{\rho}s\alpha^2]\\
\ge& 1 - \frac{n\max_{\boldsymbol{\mu}\neq\boldsymbol{\mu}'}\mathrm{KL}(\mathbb{P}_{\boldsymbol{\mu}} \parallel \mathbb{P}_{\boldsymbol{\mu}'}) + \log 2}{\log|M|}\\
\ge& m1 - \frac{ns\alpha^2 + \log 2}{\log|M|} = \Omega(1), \qquad (12)
\end{aligned}
$$

where $\mathbb{P}_{\boldsymbol{\mu}}$ is the joint distribution of $\boldsymbol{X}, Y$ with parameter $\boldsymbol{\mu}$, and the probability is over the uniform distribution of $\boldsymbol{\mu}$ on M. Here, we used $\mathrm{KL}(\mathbb{P}_{\boldsymbol{\mu}} \parallel \mathbb{P}_{\boldsymbol{\mu}'}) = \frac{1}{2}\|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_2^2$. Finally, we have

$$
\begin{aligned}
&\max_{\boldsymbol{\mu}}\mathbb{E}_{\boldsymbol{\mu}}[\mathcal{E}_{\boldsymbol{\mu}}(\mathsf{C})] \ge \tfrac{1}{|M|}\textstyle\sum_{\boldsymbol{\mu}\in M}\mathbb{E}_{\boldsymbol{\mu}}[\mathcal{E}_{\boldsymbol{\mu}}(\mathsf{C})]\\
\gtrsim&\frac{e^{-\rho^2/8}}{\rho}s\alpha^2\Pr[\mathcal{E}_{\boldsymbol{\mu}}(\mathsf{C}) \ge 4 \times 10^{-9}\frac{e^{-\rho^2/8}}{\rho}s\alpha^2]\\
\gtrsim&\tfrac{e^{-\rho^2/8}}{\rho}s\alpha^2 \gtrsim e^{-\rho^2/8}\min\{\tfrac{1}{\rho}\tfrac{p}{n}, \rho\}. \qquad (13)
\end{aligned}
$$

The next theorem states that when there are not enough samples, the best that any classifier can do is close to random guessing. Random guessing uniformly randomly assigns a sample into one of the two classes, and its classification error is $1/2$.

Theorem 3 (Impossibility of classification)**.** *For sufficiently large* p *and* n*, let* $\rho > 0$*. When* $n\max\{\rho^2, \rho^4\}/p \to 0$ *and* $p, n \to +\infty$*, then for any classifier* C *trained using* n *samples, we have*

$$
\inf_{\mathsf{C}}\max_{\|\boldsymbol{\mu}\|_2=\rho/2}\mathbb{E}_{\boldsymbol{\mu}}[\mathcal{R}_{\boldsymbol{\mu}}(\mathsf{C})] \to \tfrac{1}{2}. \qquad (14)
$$

Indeed, notice that $\max_{\boldsymbol{\mu}}\mathbb{E}_{\boldsymbol{\mu}}[R_{\boldsymbol{\mu}}(\mathsf{C})] \ge \mathbb{E}[\mathbb{E}_{\boldsymbol{\mu}}[R_{\boldsymbol{\mu}}(\mathsf{C})]]$, where the outer expectation is taken over the uniform distribution of $\boldsymbol{\mu}$ on $\|\boldsymbol{\mu}\|_2 = \rho/2$, and RHS is equivalent to a Bayesian problem with a uniform prior over $\boldsymbol{\mu}$. In the Bayesian problem, it is easy to see that any classifier's classification error is lower bounded by that of the MAP classifier. The MAP classifier $\mathsf{C}_{\mathrm{MAP}}$ is a linear classifier, and it can be written as

$$
\begin{aligned}
\mathsf{C}_{\mathrm{MAP}}(\boldsymbol{x}) &= \arg\max_y \Pr[Y = y \mid \boldsymbol{X} = \boldsymbol{x}, (\boldsymbol{x}_1, y_1), \dots]\\
&= \mathbf{1}((\textstyle\sum(2y_i - 1)\boldsymbol{x}_i^T)\boldsymbol{x} > 0). \qquad (15)
\end{aligned}
$$

Thus, we can use (4) to compute its classification error. Let $\hat{\boldsymbol{w}} = \frac{1}{n}\sum(2y_i - 1)\boldsymbol{x}_i$. When conditioned on $\boldsymbol{\mu}$, $\boldsymbol{E}_{\boldsymbol{\mu}}[\mathcal{R}(\mathsf{C}_{\mathrm{MAP}})] \to 1/2$ uniformly for all $\boldsymbol{\mu}$, because $|\boldsymbol{\mu}^T\hat{\boldsymbol{w}}/\|\hat{\boldsymbol{w}}\|_2| \to 0$ uniformly. This establishes that the MAP classifier's classification error approaches $1/2$ as $n\max\{\rho^2, \rho^4\}/p \to 0$ and $p, n \to +\infty$.

3.1.2 Upper Bound

We show that Fisher's linear discriminant achieves the minimax rate. Here we consider isotropic Gaussian classification problems with $\Pr[Y = 1] = \Pr[Y = 0] = 1/2$ and $\boldsymbol{X}|Y \sim \mathcal{N}(\boldsymbol{\mu}_Y, \sigma^2\mathbf{I})$, where $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$, and σ are unknown. We assume that $\|\boldsymbol{\mu}_1\|_2 = O(1)$, $\|\boldsymbol{\mu}_0\|_2 = O(1)$, and $\sigma = \Theta(1)$. Thus $\rho = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)} = O(1)$ and $e^{-\rho^2/8} = \Omega(1)$.

For $y = 1, 0$, let $n_y = \sum\mathbf{1}(y_i = y)$. We estimate each class's mean $\hat{\boldsymbol{\mu}}_y = \frac{1}{n_y}\sum_{y_i=y}\boldsymbol{x}_i$. And the trained classifier is $\hat{\mathsf{C}}(\boldsymbol{x}) = \mathbf{1}(\hat{\boldsymbol{w}}^T\boldsymbol{x} + \hat{b} > 0)$ where $\hat{\boldsymbol{w}} = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0$, and $\hat{b} = \frac{1}{2}(-\hat{\boldsymbol{\mu}}_1^T\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_0^T\hat{\boldsymbol{\mu}}_0)$.

Theorem 4 (Fisher's linear discriminant upper bound)**.** *Let* $\rho = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)} = O(1)$*. When* $n \gtrsim p\log\delta^{-1}/\rho^4$*, with probability at least* $1 - \delta$*, for Fisher's linear discriminant, we have*

$$
\mathcal{E}(\hat{\mathsf{C}}) \lesssim \tfrac{1}{\rho}\tfrac{p}{n}\log\tfrac{1}{\delta}. \qquad (16)
$$

The proof of Theorem 4 is based on Theorem 1 of [16], which connects parameter estimation error with classification excess risk. We state this theorem here for completeness.

Theorem 5 (Classification excess risk upper bound (Theorem 1, [16]))**.** *In a Gaussian classification problem with* $\Pr[Y = 1] = \Pr[Y = 0] = 1/2$ *and* $\boldsymbol{X} \mid Y \sim \mathcal{N}(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma})$*, the excess risk of a linear classifier* $\mathsf{C}(\boldsymbol{x}) = \mathbf{1}(\boldsymbol{w}^T\boldsymbol{x} + b > 0)$ *is bounded by*

$$
\mathcal{E}(\mathsf{C}) \lesssim \rho e^{-\rho^2/8}(e_1^2 + e_0^2) + e_1^3 + e_0^3, \qquad (17)
$$

where $e_1 = |\frac{\boldsymbol{w}^T\boldsymbol{\mu}_1 + b}{\sqrt{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}}} - \rho/2|$ *and* $e_0 = |\frac{\boldsymbol{w}^T\boldsymbol{\mu}_0 + b}{\sqrt{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}}} + \rho/2|$ *with* $\rho = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}$*.*

3.2 Sparse Setting

In the sparse setting, we consider Gaussian classification problems with $\Pr[Y = 1] = \Pr[Y = 0] = 1/2$ and $\boldsymbol{X} \mid Y \sim \mathcal{N}(\boldsymbol{\mu}_Y, \sigma^2\mathbf{I})$, with the additional requirement that $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ is s-sparse ($\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|_0 = s$).

3.2.1 Lower Bound

Similar to Section 3.1.1, for the lower bound we consider the Gaussian classification problems where $\Pr[Y=1]=\Pr[Y=0]=1/2$, $\boldsymbol{X}\mid Y\sim\mathcal{N}(\boldsymbol{\mu}_Y,\mathbf{I})$ with $\boldsymbol{\mu}_1=-\boldsymbol{\mu}_0=\boldsymbol{\mu}$. The additional requirement here is that $\boldsymbol{\mu}$ is s-sparse ($\|\boldsymbol{\mu}\|_0=s$).

The following theorem, in parallel with Theorem 1, establishes a lower bound for the excess risk of sparse classification.

Theorem 6 (Sparse classification excess risk lower bound)**.** *Let $\rho>0$ be a fixed number. For sufficiently large p and n, let $1\le s<p$, then for any classifier C trained using n samples, we have*

$$\max_{\substack{\|\boldsymbol{\mu}\|_2=\rho/2\\ \|\boldsymbol{\mu}\|_0=s}}\mathbb{E}_{\boldsymbol{\mu}}[\mathcal{E}_{\boldsymbol{\mu}}(\mathsf{C})]\gtrsim e^{-\rho^2/8}\min\{\tfrac{1}{\rho}\tfrac{s\log\frac{p}{s}}{n},\rho\}. \tag{18}$$

The proof of Theorem 6 is similar to that of Theorem 1, and it is also based on Lemma 1 and Theorem 2. Theorem 6's proof is omitted for brevity.

3.2.2 Upper Bound

We show that a modified version Fisher's linear discriminant achieves the minimax rate in sparse classification. Here we consider isotropic Gaussian classification problems with $\Pr[Y=1]=\Pr[Y=0]=1/2$ and $\boldsymbol{X}|Y\sim\mathcal{N}(\boldsymbol{\mu}_Y,\sigma^2\mathbf{I})$, where $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$, and σ are unknown. In the sparse setting, $\boldsymbol{\mu}_1-\boldsymbol{\mu}_0$ is s-sparse ($\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_0\|_0=s$). We assume that $\|\boldsymbol{\mu}_1\|_2=O(1)$, $\|\boldsymbol{\mu}_0\|_2=O(1)$, and $\sigma=\Theta(1)$. Thus $\rho=\sqrt{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}=O(1)$ and $e^{-\rho^2/8}=\Omega(1)$.

To train the classifier $\mathsf{C}(\boldsymbol{x})=\hat{\boldsymbol{w}}^T\boldsymbol{x}+\hat{b}$, we split the i.i.d. samples into two parts of size $n^{(1)}\approx n^{(2)}\approx n/2$. On the first part of the data, we estimate each class's mean $\hat{\boldsymbol{\mu}}_y=\frac{1}{n_y}\sum_{y_i^{(1)}=y}\boldsymbol{x}_i^{(1)}$ where $n_y=\sum\mathbf{1}(y_i^{(1)}=y)$ for $y=1,0$. We estimate $\hat{\boldsymbol{w}}$ by solving the following ℓ_1 regularized optimization problem

$$\min_{\boldsymbol{w}}\|\boldsymbol{w}-(\hat{\boldsymbol{\mu}_1}-\hat{\boldsymbol{\mu}_0})\|_2^2+\lambda\|\boldsymbol{w}\|_1, \tag{19}$$

with $\lambda=\Theta(\sigma\sqrt{\log p/n})$. When σ is unknown, empirically cross validation is known to be effective in selecting a suitable regularization parameter λ [9]. To estimate $\hat{b}$, we use the second part of the data and $\hat{\boldsymbol{w}}$

$$\hat{b}=\tfrac{1}{2}\hat{\boldsymbol{w}}^T(\tfrac{1}{n^{(2)}}\textstyle\sum\boldsymbol{x}_i^{(2)}). \tag{20}$$

The following theorem, which is a special case of the result for parameter recovery using ℓ_1 regularized optimization [20], shows that we can successfully recover $\boldsymbol{\mu}_1-\boldsymbol{\mu}_0$ using $\hat{\boldsymbol{w}}$.

Theorem 7 (Sparse mean estimation using ℓ_1 regularized least squares)**.** *Let $\boldsymbol{\theta}^*\in\mathbb{R}^p$ be a s-sparse vector ($\|\boldsymbol{\theta}\|_0=s<p/2$). Suppose we observe $\hat{\boldsymbol{\theta}}=\boldsymbol{\theta}^*+\boldsymbol{\epsilon}$. Set $\lambda=4\|\boldsymbol{\epsilon}\|_\infty$. If we estimate $\boldsymbol{\theta}^*$ using*

$$\hat{\boldsymbol{w}}=\arg\min_w\|\boldsymbol{w}-\hat{\boldsymbol{\theta}}\|_2^2+\lambda\|\boldsymbol{w}\|_1, \tag{21}$$

then we have

$$\begin{aligned}&\|\hat{\boldsymbol{w}}-\boldsymbol{\theta}^*\|_1\lesssim s\|\boldsymbol{\epsilon}\|_\infty,\\ &\|\hat{\boldsymbol{w}}-\boldsymbol{\theta}^*\|_2\lesssim\sqrt{s}\|\boldsymbol{\epsilon}\|_\infty.\end{aligned}\tag{22}$$

A key component in the proof of Theorem 7 is that $\mathbf{I}$ satisfies the restricted eigenvalue condition, which follows from the fact that $\mathbf{I}$ satisfies the Restricted Isometry Property [23].

The next theorem, in parallel with Theorem 4, shows that sparse Fisher's linear discriminant achieves the minimax rate up to logarithmic factors.

Theorem 8 (Sparse Fisher's linear discriminant upper bound)**.** *Let $\rho=\sqrt{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}=O(1)$. When $n\gtrsim s\log p\log\delta^{-1}/\rho^4$ and $1\le s<p/2$, then with probability at least $1-\delta$, sparse Fisher's linear discriminant satisfies*

$$\mathcal{E}(\hat{\mathsf{C}})\lesssim\tfrac{1}{\rho}\tfrac{s\log p}{n}\log\tfrac{1}{\delta}. \tag{23}$$

4 Bounds for Gaussian clustering

In this section, we present lower and upper bounds for Gaussian clustering, both in the general setting without sparsity, and in the high-dimensional sparse setting where the difference between the means is sparse.

4.1 General Setting without Sparsity

4.1.1 Lower Bound

For the lower bound, we consider Gaussian clustering problems where $\boldsymbol{X}\sim\frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_1,\mathbf{I})+\frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_0,\mathbf{I})$ with $\boldsymbol{\mu}_1=-\boldsymbol{\mu}_0=\boldsymbol{\mu}$. In this case, the optimal clustering rule's error is equal to the Bayes classification error $1-\Phi(\|\boldsymbol{\mu}\|_2)$. We will index each clustering problem with $\boldsymbol{\mu}$.

Similar to Theorem 1, the following theorem provides a lower bound for Gaussian clustering, but with a different dependency on the separation between the two clusters ($\rho=\sqrt{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}$).

Theorem 9 (Clustering excess risk lower bound)**.** *Let $\rho>0$ be a fixed number. For sufficiently large p and n, and any clustering rule C trained using n samples, we have*

$$\max_{\|\boldsymbol{\mu}\|_2=\rho/2}\mathbb{E}_{\boldsymbol{\mu}}[\mathcal{E}_{\boldsymbol{\mu}}(\mathsf{C})]\gtrsim e^{-\rho^2/8}\min\{\tfrac{1}{\rho^3}\tfrac{p}{n},\rho\}. \tag{24}$$

Similar to the proof of Theorem 1, the proof of Theorem 9 is based on the following theorem for general clustering problems, which provides a "triangle inequality" in Fano's method [25] for a fixed clustering rule's excess risk in different clustering problems.

Theorem 10 (A "triangle inequality" for clustering excess risk)**.** *Let i correspond to a clustering problem where $\boldsymbol{X} \sim \pi_1 p_1^{(i)}(\boldsymbol{X}) + \pi_0 p_0^{(i)}(\boldsymbol{X})$. Denote the optimal clustering rule as $\mathsf{C}^{*(i)}(\boldsymbol{x}) = \mathbf{1}((\pi_1^{(i)}p_1^{(i)}(\boldsymbol{x}))/(\pi_0^{(i)}p_0^{(i)}(\boldsymbol{x})) > 1)$. Suppose we have another clustering problem j. Then for any fixed clustering rule C,*

$$\begin{aligned}&\mathcal{E}_i(\mathsf{C}) + \mathcal{E}_j(\mathsf{C})\\ &\geq \min\{ \int_{\mathsf{C}^{*(i)}\neq\mathsf{C}^{*(j)}} \min\{|\pi_1^{(i)}p_1^{(i)} - \pi_0^{(i)}p_0^{(i)}|,\\ &|\pi_1^{(j)}p_1^{(j)} - \pi_0^{(j)}p_0^{(j)}|\}\; d\boldsymbol{x},\\ &\int_{\mathsf{C}^{*(i)}=\mathsf{C}^{*(j)}} \min\{|\pi_1^{(i)}p_1^{(i)} - \pi_0^{(i)}p_0^{(i)}|,\\ &|\pi_1^{(j)}p_1^{(j)} - \pi_0^{(j)}p_0^{(j)}|\}\; d\boldsymbol{x}\}. \end{aligned} \tag{25}$$

Indeed, for a clustering problem where $\boldsymbol{X} \sim \pi_1 p_1(\boldsymbol{X}) + \pi_0 p_0(\boldsymbol{X})$, a clustering rule C's excess risk $\mathcal{E}(\mathsf{C}) = \min\{\int_{\mathsf{C}\neq\mathsf{C}^*} |\pi_1 p_1 - \pi_0 p_0|\; d\boldsymbol{x}, \int_{\mathsf{C}=\mathsf{C}^*} |\pi_1 p_1 - \pi_0 p_0|\; d\boldsymbol{x}\}$ where C^* is the optimal clustering rule. This follows from the definition of clustering error (7), and the Bayes classifier's classification error (Theorem 2.2, [7]).

The proof of Theorem 9 is similar to that of Theorem 1. However, in clustering we have $\mathrm{KL}(\mathbb{P}_{\boldsymbol{\mu}}, \mathbb{P}_{\boldsymbol{\mu}'}) \lesssim \rho^4(1 - \boldsymbol{\mu}^T\boldsymbol{\mu}'/(\|\boldsymbol{\mu}\|_2\|\boldsymbol{\mu}'\|_2))$ (Proposition 24, [1]). Thus, in clustering the dependency on ρ is different from classification.

Notice that, for a classification problem, we can always remove the labels, and treat it as a clustering problem. Denote the learned clustering rule as $\hat{\mathsf{C}}$, then both $\hat{\mathsf{C}}$ and $1 - \hat{\mathsf{C}}$ can be used as classifiers for the original classification problem. Thus we have the following corollary to Theorem 3, which states that, without enough samples the best any clustering rule can do is close to random guessing.

Corollary 1 (Impossibility of clustering)**.** *For sufficiently large p and n, let $\rho > 0$. When $n\max\{\rho^2, \rho^4\}/p \to 0$ and $p, n \to +\infty$, then for any clustering rule C trained using n samples, we have*

$$\inf\nolimits_{\mathsf{C}} \max\nolimits_{\|\boldsymbol{\mu}\|_2=\rho/2} \mathbb{E}_{\boldsymbol{\mu}}[\mathcal{R}_{\boldsymbol{\mu}}(\mathsf{C})] \to \tfrac{1}{2}. \tag{26}$$

4.1.2 Upper Bound

For the upper bound, we will consider the setting with $\boldsymbol{X} \sim \boldsymbol{X} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_1, \sigma^2\mathbf{I}) + \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_0, \sigma^2\mathbf{I})$ where $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$, and σ are unknown. We assume that $\|\boldsymbol{\mu}_1\| = O(1)$, $\|\boldsymbol{\mu}_0\| = O(1)$, and $\sigma = \Theta(1)$. Thus $\rho = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)} = O(1)$ and $e^{-\rho^2/8} = \Omega(1)$.

Our method for clustering is similar to that of [1]. To estimate a clustering rule, we first compute the sample mean $\hat{\boldsymbol{m}} = \frac{1}{n}\sum \boldsymbol{x}_i$, then we compute the largest eigenvalue's normalized eigenvector $\hat{\boldsymbol{v}}$ of the sample covariance matrix $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum(\boldsymbol{x}_i - \hat{\boldsymbol{m}})(\boldsymbol{x}_i - \hat{\boldsymbol{m}})^T$. Next, we use $\hat{\mathsf{C}}(\boldsymbol{x}) = \hat{\boldsymbol{v}}^T\boldsymbol{x} - \hat{\boldsymbol{v}}^T\hat{\boldsymbol{m}}$ as the estimated clustering rule.

Theorem 11 (Clustering upper bound)**.** *For sufficiently large n and p, let $\rho = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|_2/\sigma = O(1)$. When $n \gtrsim p\log(pn)/\rho^6$, the expected excess risk of $\hat{\mathsf{C}}$ satisfies*

$$\mathbb{E}[\mathcal{E}(\hat{\mathsf{C}})] \lesssim \tfrac{1}{\rho^3}\tfrac{p}{n}\log(pn). \tag{27}$$

To prove Theorem 11, we will use the following corollary on the connection between parameter estimation error and clustering error, which follows from Theorem 5 and the definition of clustering error (7).

Corollary 2 (Clustering excess risk upper bound)**.** *In a Gaussian clustering problem with $\boldsymbol{X} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$, the excess risk of a linear clustering rule $\mathsf{C}(\boldsymbol{x}) = \mathbf{1}(\boldsymbol{w}^T\boldsymbol{x} + b > 0)$ is bounded by*

$$\begin{aligned}\mathcal{E}(\mathsf{C}) \lesssim \min\{&\rho e^{-\rho^2/8}(e_1^2 + e_0^2) + e_1^3 + e_0^3,\\ &\rho e^{-\rho^2/8}(f_1^2 + f_0^2) + f_1^3 + f_0^3\}, \end{aligned} \tag{28}$$

where $e_1 = |\frac{\boldsymbol{w}^T\boldsymbol{\mu}_1 + b}{\sqrt{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}}} - \rho/2|$, $e_0 = |\frac{\boldsymbol{w}^T\boldsymbol{\mu}_0 + b}{\sqrt{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}}} + \rho/2|$ $f_1 = |\frac{\boldsymbol{w}^T\boldsymbol{\mu}_1 + b}{\sqrt{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}}} + \rho/2|$, $f_0 = |\frac{\boldsymbol{w}^T\boldsymbol{\mu}_0 + b}{\sqrt{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}}} - \rho/2|$ and $\rho = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}$.

The proof of Theorem 11 uses Proposition 6 of [1], which gives the error of using $\hat{v}$ to estimate the direction of $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$. We state it here for completeness.

Proposition 1 (Proposition 6, [1])**.** *Suppose $n > 4p$, define $\cos\beta = \boldsymbol{\Delta\mu}^T\hat{\boldsymbol{v}}/\|\boldsymbol{\Delta\mu}\|_2$. For any $0 \le \delta < (p-1)/\sqrt{e}$, if $\max\{4/\rho^2, 2/\rho\}\sqrt{\frac{\max\{d, 8\log\delta^{-1}\}}{n}} < 1/180$, then with probability at least $1 - 12\delta - 2e^{-n/20}$, we have*

$$\sin\beta \le 14\max\{4/\rho^2, 2/\rho\}\sqrt{p}\sqrt{\tfrac{10}{n}\log\tfrac{p}{\delta}\max\{1, \tfrac{10}{n}\log\tfrac{p}{\delta}\}}. \tag{29}$$

4.2 Sparse Setting

In the sparse setting, we consider Gaussian clustering problems with $\boldsymbol{X} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_1, \sigma^2\mathbf{I}) + \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_0, \sigma^2\mathbf{I})$, with the additional requirement that $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ is s-sparse ($\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|_0 = s$).

4.2.1 Lower Bound

Similar to Section 4.1.1, for the lower bound we consider the Gaussian clustering problems where $\boldsymbol{X} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_1, \sigma^2\mathbf{I}) + \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_0, \sigma^2\mathbf{I})$ with $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_0 = \boldsymbol{\mu}$. The additional requirement here is that $\boldsymbol{\mu}$ is s-sparse ($\|\boldsymbol{\mu}\|_0 = s$).

The following theorem, in parallel with Theorem 9, establishes a lower bound for the excess risk of sparse clustering.

Theorem 12 (Sparse clustering excess risk lower bound)**.** *Let $\rho > 0$ be a fixed number. For sufficiently large p and n, let $1 \leq s < p$, then for any clustering rule C trained using n samples, we have*

$$\max_{\substack{\|\boldsymbol{\mu}\|_2=\rho/2\\ \|\boldsymbol{\mu}\|_0=s}} \mathbb{E}_{\boldsymbol{\mu}}[\mathcal{E}_{\boldsymbol{\mu}}(\mathsf{C})] \gtrsim e^{-\rho^2/8} \min\{\tfrac{1}{\rho^3}\tfrac{s\log\frac{p}{s}}{n}, \rho\}. \tag{30}$$

The proof of Theorem 12 is similar to that of Theorem 9, and it is omitted for brevity.

4.2.2 Upper Bound

Here we consider isotropic Gaussian clustering problems with $\boldsymbol{X} \sim \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_1, \sigma^2\mathbf{I}) + \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_0, \sigma^2\mathbf{I})$, where $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$, and σ are unknown. In the sparse setting, $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ is s-sparse ($\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|_0 = s$). We assume that $\|\boldsymbol{\mu}_1\| = O(1)$, $\|\boldsymbol{\mu}_0\| = O(1)$, and $\sigma = \Theta(1)$. Thus $\rho = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)} = O(1)$ and $e^{-\rho^2/8} = \Omega(1)$.

Our method for sparse clustering is similar to that of [1]. Let $S = \{i : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0 \neq 0\}$ be the set of relevant features. We first construct an estimate $\hat{S}$ of the set of relevant features. Let $\hat{\boldsymbol{m}} = \frac{1}{n}\sum \boldsymbol{x}_i$ and $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum(\boldsymbol{x}_i - \hat{\boldsymbol{m}})(\boldsymbol{x}_i - \hat{\boldsymbol{m}})^T$. Let $\hat{\tau} = \frac{1+\alpha}{1-\alpha}\min_{1\leq i\leq p}\hat{\boldsymbol{\Sigma}}_{ii}$ where $\alpha = \sqrt{6\log(np)/n} + 2\log(np)/n$. Now let $\hat{S} = \{i : \hat{\boldsymbol{\Sigma}}_{ii} > \hat{\tau}\}$. Next, we compute the largest eigenvalue's normalized eigenvector $\hat{\boldsymbol{v}}$ of the sample covariance matrix restricted to coordinates in $\hat{S}$. Finally, we use $\hat{\mathsf{C}}(\boldsymbol{x}) = \hat{\boldsymbol{v}}^T\boldsymbol{x}_{\hat{S}} - \hat{\boldsymbol{v}}^T\hat{\boldsymbol{m}}_{\hat{S}}$ as the estimated clustering rule, where $\boldsymbol{x}_{\hat{S}}$ and $\hat{\boldsymbol{m}}_{\hat{S}}$ are $\boldsymbol{x}$ and $\hat{\boldsymbol{m}}$ restricted to coordinates in $\hat{S}$, respectively.

Theorem 13 (Clustering upper bound)**.** *For sufficiently large n, p, and $s < n/4$, let $\rho = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0\|_2/\sigma$. When $\alpha < 1/4$ and $n \gtrsim s^2\log(pn)/\rho^8$, the expected excess risk of $\hat{\mathsf{C}}$ satisfies*

$$\mathbb{E}[\mathcal{E}(\hat{\mathsf{C}})] \lesssim \tfrac{1}{\rho^3}\tfrac{s\log(ns)}{n} + \tfrac{s}{\rho^2}\left(\tfrac{\log(pn)}{n}\right)^{\frac{1}{2}}. \tag{31}$$

Although the above bound is not as tight as other previous bounds, we would like to note that sparse Gaussian mixture problems and sparse principal component analysis problems are both computationally and statistically challenging [3, 2, 19, 24, 10].

The proof of Theorem 13 uses Proposition 9 of [1] regarding performance of support recovery.

Proposition 2 (Proposition 9, [1])**.** *Assume that $n \geq 1$, $p \geq 2$, and $\alpha < 1/4$. Define $\tilde{S} = \{i : |(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_i| \geq 4\sigma\sqrt{\alpha}\}$. Then $\tilde{S} \subseteq \hat{S} \subseteq S$ with probability at least $1 - 6/n$.*

We briefly sketch Theorem 13's proof. Let $\hat{\beta}$ be the angle between $\hat{\boldsymbol{v}}_{\hat{S}}$ and $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$, $\bar{\beta}$ the angle between $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_{\hat{S}}$ and $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$, and β the angle between $\hat{\boldsymbol{v}}_{\hat{S}}$ and $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_{\hat{S}}$. Proposition 1 shows β is small. Proposition 2 shows $\bar{\beta}$ is small. Using the triangle inequality in spherical geometry we have $\hat{\beta} \leq \bar{\beta} + \beta$. Thus establishing Theorem 13.

4.3 Comparison of Different Clustering Error Definitions

Here we compare our definition of clustering error with that of [1], and show relationships between the two definitions. We define clustering error as

$$\min\{\Pr[\mathsf{C} \neq Y], 1 - \Pr[\mathsf{C} \neq Y]\} \tag{32}$$

with Y being the latent indicator in the Gaussian mixture model, whereas in [1] it is defined as $\min\{\Pr[\mathsf{C} \neq \mathsf{C}^*], 1 - \Pr[\mathsf{C} \neq \mathsf{C}^*]\}$.

One advantage of our definition is that we can empirically evaluate the clustering error if true labels are given. This is because our definition of clustering error is the same as classification error in the corresponding classification problem. However, empirically evaluating [1]'s clustering error is not straightforward even if true labels are available.

Proposition 3 ([1]'s clustering error upper bound (Proposition 7, [1]))**.** *For the clustering problem considered in 4.1.2, [1]'s clustering error of a linear clustering rule $\mathsf{C}(\boldsymbol{x}) = \mathbf{1}(\boldsymbol{w}^T\boldsymbol{x} + b > 0)$ is bounded by*

$$\min\{\Pr[\mathsf{C} \neq \mathsf{C}^*], 1 - \Pr[\mathsf{C} \neq \mathsf{C}^*]\} \lesssim (\epsilon_1 + \epsilon_2\rho + |\sin\beta|), \tag{33}$$

if $|b + w^T\frac{(\boldsymbol{\mu}_1+\boldsymbol{\mu}_0)}{2}| \leq (\sigma\epsilon_1 + \|\frac{\boldsymbol{\mu}_1-\boldsymbol{\mu}_0}{2}\|_2\epsilon_2)\|w\|_2$ for some $\epsilon_1 \geq 0$ and $0 \leq \epsilon_2 \leq \frac{1}{4}$, and $|\sin\beta| \leq 1/\sqrt{5}$, where $\cos\beta = \frac{w^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)}{\|w\|_2\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_0\|_2}$.

Also, our clustering error definition captures the intrinsic hardness of the clustering problem better than [1]'s definition. For [1]'s definition, a linear clustering rule's clustering error can be upper bounded using Proposition 3. Following [1], with high probability we have $\min\{\Pr[\mathsf{C} \neq \mathsf{C}^*], 1 - \Pr[\mathsf{C} \neq \mathsf{C}^*]\} \lesssim \sin\beta$. For our definition, from Corollay 2 it is easy to derive that $\min\{\Pr[\mathsf{C} \neq Y], 1 - \Pr[\mathsf{C} \neq Y]\} \lesssim \rho|\sin\beta|^2 + |\sin\beta|^3$ holds with high probability. Thus when $|\sin\beta| \lesssim \rho$, which requires $n \gg p/\rho^6$, we can easily "convert" these two definitions into one another up to a constant. Despite this observation, [1]'s definition does not, in a straightforward fashion, exhibit an impossibility result similar to Corollary 1. When $n \ll p/\rho^2$, Corollary 1 shows the clustering error defined by us approaches 1/2, which can be interpreted as random guessing. But it is not straightforward to make such an interpretation using [1]'s definition, and to our knowledge an impossibility result using [1]'s definition of clustering error is not known.

5 Conclusion

In this paper, we presented minimax lower and upper bounds for Gaussian classification and clustering. Our results explicitly show how the statistical difficulty depends on dimension p, number of samples n, sparsity of the optimal classifier/clustering rule s, and separation between two classes/clusters $\rho = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)}$. Our results here focus on classification and clustering of two isotropic (spherical) Gaussian distributions with equal proportions.

In future work, we plan to improve the bounds' dependency on separation ρ. Here we have shown that, in the general non-sparse setting when $n \max\{\rho^2, \rho^4\}/p \to 0$ and $p, n \to +\infty$, both classification and clustering are close to impossible. But efficient procedures for classification (Fisher's linear discriminant) and clustering (principal component analysis) achieve minimax rates when $n\rho^4/p \to +\infty$ and $n\rho^6/p \to +\infty$. In the sparse setting, for classification and clustering we have sample lower bounds $n = \Omega(s \log \frac{p}{s}/\rho^2)$ for classification and $n = \Omega(s \log \frac{p}{s}/\rho^4)$ for clustering. But efficient procedures for classification (sparse Fisher's linear discriminant) and clustering (sparse principal component analysis) requires $n\rho^4/(s \log p) \to +\infty$ for classification and $n\rho^8/(s \log p) \to +\infty$ for clustering. We conjecture that there is a statistical and computational tradeoff in both classification and clustering problems, similar to the result on sparse principal component detection [2]. Another part of future work is to extend this paper's results to general Gaussian distributions.

6 Acknowledgments

C.C. acknowledges partial support of NSF Grants 1609279, 1056028 and 1302435, and the U.S. DoT through the Data-Supported Transportation Operations and Planning (D-STOP) Tier 1 University Transportation Center. P.R. acknowledges the support of ARO via W911NF-12-1-0390 and NSF via IIS-1149803, IIS-1320894, IIS-1447574, and DMS-1264033, and NIH via R01 GM117594-01 as part of the Joint DMS/NIGMS Initiative to Support Research at the Interface of the Biological and Mathematical Sciences.

References

[1] M. Azizyan, A. Singh, and L. Wasserman. Minimax Theory for High Dimensional Gaussian Mixtures with Sparse Mean Separation. In *Advances in Neural Information Processing Systems*, pages 2139–2147, 2013.

[2] Q. Berthet and P. Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory*, pages 1046–1066, 2013.

[3] Q. Berthet and P. Rigollet. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013.

[4] P. J. Bickel and E. Levina. Some theory for Fisher's linear discriminant function,'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, pages 989–1010, 2004.

[5] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence.* Oxford university press, 2013.

[6] T. Cai and W. Liu. A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association*, 2012.

[7] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition.* Springer Science & Business Media, 2013.

[8] J. Fan and Y. Fan. High dimensional classification using features annealed independence rules. *Annals of Statistics*, 36(6):2605, 2008.

[9] C. Giraud. *Introduction to high-dimensional statistics*, volume 138. CRC Press, 2014.

[10] M. Hardt and E. Price. Tight bounds for learning a mixture of two Gaussians. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 753–760. ACM, 2015.

[11] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics. Springer New York, 2013.

[12] S. Jukna. *Extremal combinatorics: with applications in computer science.* Springer Science & Business Media, 2011.

[13] A. Krause, P. Perona, and R. G. Gomes. Discriminative clustering by regularized information maximization. In *Advances in Neural Information Processing Systems*, pages 775–783, 2010.

[14] G. Lecué. Optimal rates of aggregation in classification under low noise assumption. *Bernoulli*, 13(4):1000–1022, 2007.

[15] G. Lecué. Classification with minimax fast rates for classes of Bayes rules with sparse representation. *Electronic Journal of Statistics*, 2:741–773, 2008.

[16] T. Li, A. Prasad, and P. Ravikumar. Fast Classification Rates for High-dimensional Gaussian Generative Models. In *Advances in Neural Information Processing Systems*, pages 1054–1062, 2015.

[17] E. Mammen and A. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

[18] P. Massart. *Concentration inequalities and model selection*. Springer, 2007.

[19] A. Moitra and G. Valiant. Settling the Polynomial Learnability of Mixtures of Gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 93–102. IEEE, 2010.

[20] S. Negahban, B. Yu, M. Wainwright, and P. Ravikumar. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

[21] J. Shao, Y. Wang, X. Deng, and S. Wang. Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of statistics*, 39(2):1241–1265, 2011.

[22] Y. Shi and F. Sha. Information-Theoretical Learning of Discriminative Clusters for Unsupervised Domain Adaptation. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1079–1086, 2012.

[23] S. A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[24] N. Verzelen and E. Arias-Castro. Detection and feature selection in sparse mixture models. *arXiv preprint arXiv:1405.1478*, 2014.

[25] B. Yu. Assouad, Fano, and Le Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

Detailed Proofs

A Classification

A.1 General Setting without Sparsity

A.1.1 Lower Bounds

Proof of Theorem 1 (Classification excess risk lower bound). For simplicity, let $\omega = \rho/2$. We will consider the following set of parameters

$$M = \{\boldsymbol{\mu} \mid (\sqrt{\omega^2 - s\alpha^2}), \pm\alpha\psi_1, \dots, \pm\alpha\psi_{p-1}) \in \mathbb{R}^p\} \tag{34}$$

where $s = (p-1)/6 > p/8$, and $\psi_1, \dots, \psi_{p-1}$ are as given in Lemma 1. And we set

$$\alpha = 0.001 \min\{\sqrt{\frac{1}{n}}, \frac{\omega}{\sqrt{s}}\} \tag{35}$$

From Lemma 1, we have

$$\log |M| \geq \frac{s}{5} \log \frac{p-1}{s} \geq 0.04p \tag{36}$$

Let $\mathbb{P}_{\boldsymbol{\mu}}$ denote the joint distribution of $\boldsymbol{X}, Y$ with $\Pr[Y=1] = \Pr[Y=0] = 1/2$, $\boldsymbol{X} \mid Y = 1 \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$, and $\boldsymbol{X} \mid Y = 0 \sim \mathcal{N}(-\boldsymbol{\mu}, \mathbf{I})$.

For any $\boldsymbol{\mu}$, $\boldsymbol{\mu}'$, we have

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_{\boldsymbol{\mu}} \parallel \mathbb{P}_{\boldsymbol{\mu}'}) &= \mathbb{E}_{\boldsymbol{\mu}}[\log \frac{\mathbb{P}_{\boldsymbol{\mu}}}{\mathbb{P}_{\boldsymbol{\mu}'}}] \\
&= \frac{1}{2}\mathbb{E}_{\mathcal{N}(\boldsymbol{\mu},\mathbf{I})}[\log \frac{\mathcal{N}(\boldsymbol{\mu},\mathbf{I})}{\mathcal{N}(\boldsymbol{\mu}',\mathbf{I})}] + \frac{1}{2}\mathbb{E}_{\mathcal{N}(-\boldsymbol{\mu},\mathbf{I})}[\log \frac{\mathcal{N}(-\boldsymbol{\mu},\mathbf{I})}{\mathcal{N}(-\boldsymbol{\mu}',\mathbf{I})}] \\
&= \frac{1}{2}\|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_2^2
\end{aligned} \tag{37}$$

Consider any $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathbb{R}^p$ with $\|\boldsymbol{\mu}\|_2 = \|\boldsymbol{\mu}'\|_2 = \omega$, for any fixed classifier C, Theorem 2 gives

$$\begin{aligned}
\mathcal{E}_{\boldsymbol{\mu}}(\mathsf{C}) + \mathcal{E}_{\boldsymbol{\mu}'}(\mathsf{C}) &\geq \frac{1}{2}\int_{\mathsf{C}^*_{\boldsymbol{\mu}} \neq \mathsf{C}^*_{\boldsymbol{\mu}'}} \min\{|\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}) - \mathcal{N}(-\boldsymbol{\mu}, \mathbf{I})|, |\mathcal{N}(\boldsymbol{\mu}', \mathbf{I}) - \mathcal{N}(-\boldsymbol{\mu}', \mathbf{I})|\} \, d\boldsymbol{x} \\
&\overset{\text{by symmetry}}{=} 2 \iint_{\substack{u \geq 0 \\ 0 \leq v \leq (\tan\frac{\phi}{2})u}} \mathcal{N}(\begin{bmatrix} 0 \\ \omega \end{bmatrix}, \mathbf{I})(\begin{bmatrix} u \\ v \end{bmatrix}) - \mathcal{N}(\begin{bmatrix} 0 \\ -\omega \end{bmatrix}, \mathbf{I}))(\begin{bmatrix} u \\ v \end{bmatrix}) \, du \, dv \\
&= \frac{1}{\pi} \iint_{\substack{u \geq 0 \\ 0 \leq v \leq (\tan\frac{\phi}{2})u}} \exp(-\frac{1}{2}u^2 - \frac{1}{2}(v-\omega)^2) - \exp(-\frac{1}{2}u^2 - \frac{1}{2}(v+\omega)^2) \, du \, dv \\
&= \frac{e^{-\omega^2/2}}{\pi} \iint_{\substack{u \geq 0 \\ 0 \leq v \leq (\tan\frac{\phi}{2})u}} \exp(-\frac{1}{2}u^2 - \frac{1}{2}v^2)(e^{\omega v} - e^{-\omega v}) \, du \, dv \\
&\overset{\text{because } e^{\omega v} - e^{-\omega v} \geq 2\omega v \text{ for } v \geq 0}{\geq} \frac{\omega e^{-\omega^2/2}}{\pi} \iint_{\substack{u \geq 0 \\ 0 \leq v \leq (\tan\frac{\phi}{2})u}} v e^{-(u^2+v^2)/2} \, du \, dv \\
&= \frac{1}{\sqrt{2\pi}}(1 - \cos\frac{\phi}{2})\omega e^{-\omega^2/2} \\
&= \sqrt{\frac{2}{\pi}} \omega e^{-\omega^2/2} \sin^2\frac{\phi}{4} \qquad (38)
\end{aligned}$$

where $\phi = \arccos \frac{\boldsymbol{\mu}^T\boldsymbol{\mu}'}{\omega^2} \leq \frac{\pi}{2}$.
For any $\boldsymbol{\mu} \neq \boldsymbol{\mu}' \in M$, we have

$$\begin{aligned}
\sin^2\frac{\phi}{4} &= 1 - \cos^2\frac{\phi}{4} \\
&\geq \frac{1}{16}(1 - \cos^2\phi) \\
&= \frac{1}{16}(1 - \frac{\boldsymbol{\mu}^T\boldsymbol{\mu}'}{\omega^2}) \\
&\geq 4 \times 10^{-9}\frac{1}{\omega^2}s\alpha^2 \qquad (39)
\end{aligned}$$

Next, notice that

$$\begin{aligned}
\max_{\boldsymbol{\mu}} \mathbb{E}_{\boldsymbol{\mu}}[\mathcal{E}_{\boldsymbol{\mu}}(\mathsf{C})] &\geq \frac{1}{|M|}\sum_{\boldsymbol{\mu} \in M} \mathbb{E}_{\boldsymbol{\mu}}[\mathcal{E}_{\boldsymbol{\mu}}(\mathsf{C})] \\
&\overset{\text{by Markov's inequality}}{\gtrsim} \omega e^{-\omega^2/2} s\alpha^2 \Pr[\mathcal{E}_{\boldsymbol{\mu}}(\mathsf{C}) \geq 2 \times 10^{-9}\frac{e^{-\omega^2/2}}{\omega}s\alpha^2] \qquad \boldsymbol{\mu} \text{ is uniformly randomly chosen from } M \qquad (40)
\end{aligned}$$

For any fixed $\boldsymbol{\mu}$, using (38), we have

$$\mathcal{E}_{\boldsymbol{\mu}} > 2 \times 10^{-9}\frac{e^{-\omega^2/2}}{\omega}s\alpha^2 \Rightarrow \mathcal{E}_{\boldsymbol{\mu}'} < 2 \times 10^{-9}\frac{e^{-\omega^2/2}}{\omega}s\alpha^2 \qquad (41)$$

for all other $\boldsymbol{\mu}' \neq \boldsymbol{\mu}$. This observation reduces the problem to a testing problem.

Thus, we can use Fano's theorem [25] to lower bound

$$\begin{aligned}\Pr[\mathcal{E}_{\boldsymbol{\mu}}(\mathsf{C}) \geq 2 \times 10^{-9} \frac{e^{-\omega^2/2}}{\omega} s\alpha^2] &\geq 1 - \frac{n \max_{\boldsymbol{\mu} \neq \boldsymbol{\mu}'} \mathrm{KL}(\mathbb{P}_{\boldsymbol{\mu}} \parallel \mathbb{P}_{\boldsymbol{\mu}'}) + \log 2}{\log |M|} \\ &\geq 1 - \frac{ns\alpha^2 + \log 2}{\log |M|} = \Omega(1) \end{aligned} \tag{42}$$

Combining (40) and (42), we see that for any classifier C

$$\max_{\|\boldsymbol{\mu}\|_2 = \rho/2} \mathbb{E}_{\boldsymbol{\mu}}[\mathcal{E}_{\boldsymbol{\mu}}(\mathsf{C})] \gtrsim e^{-\rho^2/8} \min\{\frac{1}{\rho}\frac{p}{n}, \rho\} \tag{43}$$

□

Proof of Theorem 3 (Impossibility of classification). Indeed, notice that $\max_{\boldsymbol{\mu}} \mathbb{E}_{\boldsymbol{\mu}}[R_{\boldsymbol{\mu}}(\mathsf{C})] \geq \mathbb{E}[\mathbb{E}_{\boldsymbol{\mu}}[R_{\boldsymbol{\mu}}(\mathsf{C})]]$, where the outer expectation is taken over the uniform distribution of $\boldsymbol{\mu}$ on $\|\boldsymbol{\mu}\|_2 = \rho/2$, and RHS is equivalent to a Bayesian problem with a uniform prior over $\boldsymbol{\mu}$. In the Bayesian problem, it is easy to see that any classifier's classification error is lower bounded by that of the MAP classifier.

The MAP classifier $\mathsf{C}_{\mathrm{MAP}}$ can be written as

$$\begin{aligned}\mathsf{C}_{\mathrm{MAP}}(\boldsymbol{x}) &= \arg\max_y \Pr[Y = y \mid \boldsymbol{X} = \boldsymbol{x}, (\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)] \\ &= \arg\max_y \Pr[Y = y, \boldsymbol{X} = \boldsymbol{x}, (\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)]\end{aligned} \tag{44}$$

For simplicity, write $\boldsymbol{X}_0 = \boldsymbol{x}$, $Y_0 = y$. Also denote $v = 2y - 1$, $\omega = \rho/2$. Then the quantity in (44) can be written as

$$\begin{aligned}\mathbb{E}[\prod p(\boldsymbol{x}_i, v_i|\boldsymbol{\mu})] &\propto \mathbb{E}[\prod \exp(-\frac{1}{2}\|\boldsymbol{x}_i - v_i\boldsymbol{\mu}\|_2^2)] \\ &\propto \mathbb{E}[\exp(\boldsymbol{\mu}^T \sum v_i \boldsymbol{x}_i)]\end{aligned} \tag{45}$$

where the expectation is taken over the uniform distribution of $\boldsymbol{\mu}$ on a sphere of radius ω around the origin.

For $\boldsymbol{a} \in \mathbb{R}^p$, the expectation of $\exp(\boldsymbol{a}^T\boldsymbol{\mu})$ over the uniform distribution of $\boldsymbol{\mu}$ on a sphere of fixed radius ω around the origin only depends on $\|\boldsymbol{a}\|_2$. This is because $\boldsymbol{a}^T\boldsymbol{\mu} = \|\boldsymbol{a}\|_2\|\boldsymbol{\mu}\|_2 \cos\phi(\boldsymbol{a}, \boldsymbol{\mu})$ where $\phi(\boldsymbol{a}, \boldsymbol{\mu})$ is the angle between the two vectors $\boldsymbol{a}$ and $\boldsymbol{\mu}$, and the expectation is now uniformly over all angles $\cos\phi(\boldsymbol{a}, \boldsymbol{\mu})$. Thus the expectation remains constant as long as $\|\boldsymbol{a}\|_2$ is constant.

Because $e^x + e^{-x}$ is monotonically increasing on $[0, +\infty)$, we see that the expectation is monotonically increasing in $\|\boldsymbol{a}\|_2$.

Thus we can write

$$\begin{aligned}\mathsf{C}_{\mathrm{MAP}}(\boldsymbol{x}) &= \arg\max_y \|(2y-1)\boldsymbol{x} + \sum (2y_i - 1)\boldsymbol{x}_i\|_2 \\ &= \mathbf{1}((\sum (2y_i - 1)\boldsymbol{x}_i^T)\boldsymbol{x} > 0)\end{aligned} \tag{46}$$

This is a linear classifier, so we can use (4) compute its classification error. Let $\hat{\boldsymbol{w}} = \frac{1}{n}\sum(2Y_i - 1)\boldsymbol{X}_i$, and $\boldsymbol{\Delta} = \hat{\boldsymbol{w}} - \boldsymbol{\mu}$. Notice that $\boldsymbol{\Delta} \sim \mathcal{N}(0, \frac{1}{n}\mathbf{I})$. For simplicity write $\omega = \rho/2$, $\epsilon = n\max\{\omega^4, \omega^2\}/p$, and note that $n\max\{\rho^2, \rho^4\}/p \to 0$ is equivalent to $n\max\{\omega^4, \omega^2\}/p = \epsilon \to 0$. Conditioned on $\boldsymbol{\mu}$,

$$\begin{aligned}|\frac{\boldsymbol{\mu}^T\hat{\boldsymbol{w}}}{\|\hat{\boldsymbol{w}}\|_2}| &= |\frac{\boldsymbol{\mu}^T\boldsymbol{\Delta} + \omega^2}{\sqrt{\omega^2 + 2\boldsymbol{\mu}^T\boldsymbol{\Delta} + \|\boldsymbol{\Delta}\|_2^2}}| \\ &\leq \frac{|\boldsymbol{\mu}^T\boldsymbol{\Delta}|}{\sqrt{\omega^2 + 2\boldsymbol{\mu}^T\boldsymbol{\Delta} + \|\boldsymbol{\Delta}\|_2^2}} + \frac{\omega^2}{\sqrt{\omega^2 + 2\boldsymbol{\mu}^T\boldsymbol{\Delta} + \|\boldsymbol{\Delta}\|_2^2}}\end{aligned} \tag{47}$$

where we have ignored 0 probability events where the denominator is 0. Notice that, by Bernstein's inequality, with probability at least $1-\delta$, $|\boldsymbol{\mu}^T\boldsymbol{\Delta}| \lesssim \omega\sqrt{\frac{\log\delta^{-1}}{n}}$ and $\frac{p}{n}(1+\Theta(\sqrt{\frac{\log\delta^{-1}}{n}})) \geq \|\boldsymbol{\Delta}\|_2^2 \geq \frac{p}{n}(1-\Theta(\sqrt{\frac{\log\delta^{-1}}{n}}))$ simultaneously hold. So for sufficiently large n, letting $a = \sqrt{\frac{\log\delta^{-1}}{n}}$ and $\delta = 1/n \to 0$, (47) is bounded by

$$\begin{aligned} O(\frac{\max\{\omega,\omega^2\}}{\frac{p}{n}-\omega}) &\lesssim \frac{1}{\frac{p}{n\max\{\omega^4,\omega^2\}}-1} \\ &\lesssim \epsilon \to 0 \end{aligned} \tag{48}$$

Thus conditioned on $\boldsymbol{\mu}$, by Taylor's theorem we have $\mathbb{E}_{\boldsymbol{\mu}}[\mathcal{R}(\mathsf{C}_{\mathrm{MAP}})] \geq (\frac{1}{2} - O(\sqrt{\epsilon}))(1-O(1/n))$. Notice that this bound is uniform over $\|\boldsymbol{\mu}\|_2 = \omega$, thus we have

$$\inf_{\mathsf{C}} \max_{\boldsymbol{\mu}} \mathbb{E}_{\boldsymbol{\mu}}[\mathcal{R}_{\boldsymbol{\mu}}(\mathsf{C})] \geq \mathbb{E}[\mathbb{E}_{\boldsymbol{\mu}}[\mathcal{R}_{\boldsymbol{\mu}}(\mathsf{C})]] \geq (\frac{1}{2} - O(\sqrt{\epsilon}))(1-O(1/n)) \to \frac{1}{2} \tag{49}$$

□

A.1.2 Upper Bound

Proof of Theorem 4 (Fisher's linear discriminant upper bound). To use Theorem 5, we will bound

$$\left| \frac{\boldsymbol{\mu}_1^T \hat{w} + \hat{b}}{\sigma\|\hat{\boldsymbol{w}}\|_2} - \frac{\rho}{2} \right| \tag{50}$$

Let $\boldsymbol{w}^* = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$, and $b^* = \frac{1}{2}(-\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0^T\boldsymbol{\mu}_0)$. Notice that $\|\boldsymbol{w}^*\|_2 = \rho\sigma$. And denote $\boldsymbol{\Delta} = \hat{\boldsymbol{w}} - \boldsymbol{w}^*$, and we have $\mathbb{E}[\boldsymbol{\Delta}] = 0$.

With probability at least $1-\delta$, for $y = 0, 1$ we simultaneously have $|n_y/n - 1/2| \lesssim \sqrt{\log\delta^{-1}/n}$, thus we also simultaneously have $\|\boldsymbol{\mu}_y - \boldsymbol{\mu}_y\|_2 \lesssim \sigma\sqrt{\frac{p}{n}\log\delta^{-1}}$.

Thus $|\|\hat{\boldsymbol{w}}\|_2 - \sigma\rho| \lesssim \frac{|\boldsymbol{w}^{*T}\boldsymbol{\Delta}| + \|\boldsymbol{\Delta}\|_2^2}{\rho\sigma} \lesssim \sigma\sqrt{\frac{p}{n}\log\delta^{-1}}$, where we used the fact that $n \gtrsim p\log\delta^{-1}/\rho^4 \gtrsim p\log\delta^{-1}/\rho^2$.

And from Cauchy-Schwarz inequality and $\|\boldsymbol{\mu}_1\|_2 = O(1)$, we also have $|\boldsymbol{\mu}_1^T(\hat{\boldsymbol{w}} - \boldsymbol{w}^*)| \lesssim \sigma\sqrt{\frac{p}{n}\log\delta^{-1}}$.

Next, we have

$$\begin{aligned} |\hat{b} - b^*| &\lesssim \sum_y \|\hat{\boldsymbol{\mu}}_y - \boldsymbol{\mu}_y\|_2 \\ &\lesssim \sigma\sqrt{\frac{p}{n}\log\delta^{-1}} \end{aligned} \tag{51}$$

which also follows from Cauchy-Schwarz inequality and the boundedness of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$.

Thus, from the boundedness of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_0$, and σ, we have

$$\begin{aligned} \left| \frac{\boldsymbol{\mu}_1^T \hat{w} + \hat{b}}{\sigma\|\hat{\boldsymbol{w}}\|_2} - \frac{\rho}{2} \right| &= \frac{1}{\sigma}\left| \frac{\boldsymbol{\mu}_1^T \hat{w} + \hat{b}}{\|\hat{\boldsymbol{w}}\|_2} - \frac{\sigma\rho}{2} \right| \\ &\lesssim \frac{1}{\sigma}\frac{(\rho\sigma + (\rho\sigma)^2)\sigma\sqrt{\frac{p}{n}\log\delta^{-1}}}{(\rho\sigma)^2} \\ &\lesssim \frac{1}{\rho}\sqrt{\frac{p}{n}\log\delta^{-1}} \end{aligned} \tag{52}$$

Similarly, we can show that $\left| \frac{\boldsymbol{\mu}_0^T \hat{w} + \hat{b}}{\sigma\|\hat{\boldsymbol{w}}\|_2} + \frac{\rho}{2} \right| \lesssim \frac{1}{\rho}\sqrt{\frac{p}{n}\log\delta^{-1}}$.

Finally, from Theorem 5, and using the fact that when $n \gtrsim p\log\delta^{-1}/\rho^4$ we have $\sum|e_y - \rho/2|^3 \lesssim \rho\sum|e_y - \rho/2|^2$, we conclude that

$$\mathcal{E}(\hat{\mathsf{C}}) \lesssim \frac{1}{\rho}\frac{p}{n}\log\frac{1}{\delta} \tag{53}$$

□

A.2 Sparse Setting

Proof of Theorem 7 (Sparse mean estimation using ℓ_1 regularized least squares). Notice that $\mathbf{I}$ satisfies the Restricted Isometry Property with for any $s < p/2$, from [23] we know that it also satisfies the restricted eigenvalue conditions with some absolute constant $\Omega(1)$ and sparsity $2s < p$.

The proof here is a special case of Theorem 1 in [20].

First, notice that

$$\|\hat{\boldsymbol{w}} - \hat{\boldsymbol{\theta}}\|_2^2 + \lambda\|\hat{\boldsymbol{w}}\|_1 \leq \|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}\|_2^2 + \lambda\|\boldsymbol{\theta}^*\|_1 \tag{54}$$

Denote $\boldsymbol{\Delta} = \hat{\boldsymbol{w}} - \boldsymbol{\theta}^*$

When $\lambda = 2\|\boldsymbol{\epsilon}\|_\infty$, we have

$$-2\|\boldsymbol{\epsilon}\|_\infty\|\hat{\boldsymbol{w}} - \boldsymbol{\theta}^*\|_1 \leq -|2\boldsymbol{\epsilon}^T(\hat{\boldsymbol{w}} - \boldsymbol{\theta}^*)| \leq \lambda(\|\boldsymbol{\theta}^*\|_1 - \|\hat{\boldsymbol{w}}\|_1) \leq 3\|\boldsymbol{\epsilon}\|_\infty(\|\boldsymbol{\theta}^*\|_1 - \|\hat{\boldsymbol{w}}\|_1) \tag{55}$$

which implies

$$3\|\boldsymbol{\Delta}_{\bar{S}}\|_1 \leq \|\boldsymbol{\Delta}_S\|_1 \tag{56}$$

where $S = \{i : \boldsymbol{\theta}_i^* \neq 0\}$ and has s elements.

From (54), we have

$$\begin{aligned} &\|\boldsymbol{\Delta} - \boldsymbol{\epsilon}\|_2 \leq \lambda\|\boldsymbol{\Delta}\|_1 + \|\boldsymbol{\epsilon}\|_2^2 \\ \Leftrightarrow &\|\boldsymbol{\Delta}\|_2^2 \lesssim \|\boldsymbol{\epsilon}\|_\infty\|\boldsymbol{\Delta}\|_1 \\ \Leftrightarrow &\frac{1}{s}\|\boldsymbol{\Delta}\|_1^2 \lesssim \|\boldsymbol{\epsilon}\|_\infty\|\boldsymbol{\Delta}\|_1 \\ \Leftrightarrow &\|\boldsymbol{\Delta}\|_1 \lesssim s\|\boldsymbol{\epsilon}\|_\infty \end{aligned} \tag{57}$$

where we applied the restricted eigenvalue condition.

From (57), we also have

$$\|\boldsymbol{\Delta}\|_2 \lesssim \sqrt{\|\boldsymbol{\epsilon}\|_\infty\|\boldsymbol{\Delta}\|_1} \lesssim \sqrt{s}\|\boldsymbol{\epsilon}\|_\infty \tag{58}$$

□

Proof of Theorem 8 (Sparse Fisher's linear discriminant upper bound). Let $\boldsymbol{w}^* = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$, and $b^* = \frac{1}{2}(-\boldsymbol{\mu}_1^T\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0^T\boldsymbol{\mu}_0)$.

Write $\boldsymbol{\epsilon} = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0 - \boldsymbol{w}^*$, then we have $\mathbb{E}[\boldsymbol{\epsilon}] = 0$, and each $\boldsymbol{\epsilon}_i$ is a sub-Gaussian random variable with parameter $O(\sigma^2/n)$. Thus, from a well known result on the maximum of p sub-Gaussian random variables, we know that, with probability at least $1 - \delta$

$$\|\boldsymbol{\epsilon}\|_\infty \lesssim \sigma\sqrt{\frac{\log p}{n}\log\delta^{-1}} \tag{59}$$

Conditioned on $\hat{\boldsymbol{w}}$, we can see that $\hat{b}$ is a Gaussian random variable with variance $O(\sigma^2/n)$, thus with high probability we have

$$|\hat{b} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)^T\hat{\boldsymbol{w}}| \lesssim \sigma\sqrt{\frac{s\log p}{n}\log\delta^{-1}} \tag{60}$$

After applying Theorem 7, we have, with probability at least $1 - \delta$

$$\begin{aligned} |\hat{b} - b^*| &\lesssim \sigma\sqrt{\frac{s\log p}{n}\log\delta^{-1}} \\ \|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2 &\lesssim \sigma\sqrt{\frac{s\log p}{n}\log\delta^{-1}} \end{aligned} \tag{61}$$

The rest of the proof is similar to that of Theorem 4, and it is omitted for brevity.

□

B Clustering

B.1 General Setting without Sparsity

B.1.1 Lower Bound

Proof of Theorem 9 (Clustering excess risk lower bound). In clustering, we have $\mathrm{KL}(\mathbb{P}_{\boldsymbol{\mu}}, \mathbb{P}_{\boldsymbol{\mu}'}) \lesssim \rho^4(1 - \boldsymbol{\mu}^T\boldsymbol{\mu}'/(\|\boldsymbol{\mu}\|_2\|\boldsymbol{\mu}'\|_2))$ (Proposition 24, [1]).

The proof is similar to that of Theorem 1.

For simplicity, let $\omega = \rho/2$. We will consider the following set of parameters

$$M = \{\boldsymbol{\mu} \mid (\sqrt{\omega^2 - s\alpha^2}), \pm\alpha\psi_1, \dots, \pm\alpha\psi_{p-1}) \in \mathbb{R}^p\} \tag{62}$$

where $s = (p-1)/6 > p/8$, and $\psi_1, \dots, \psi_{p-1}$ are as given in Lemma 1. And we set

$$\alpha = 0.001 \min\{\frac{1}{\omega\sqrt{n}}, \frac{\omega}{\sqrt{s}}\} \tag{63}$$

The rest of the proof is omitted for brevity.

□

B.1.2 Upper Bound

Proof of Theorem 11 (Clustering upper bound). Our proof is based on results from [1].

First, note that, with probability at least $1-\delta$, we have $\|\boldsymbol{m} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)\|_2 \lesssim \sigma\sqrt{\frac{p}{n}\log\delta^{-1}}$.

Next, we will use Proposition 6 of [1] (Propositions 1).

Denote $\boldsymbol{\Delta\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$.

Thus, when $\delta = \frac{1}{n}$, it is easy to see that, with probability at lest $1-\delta$,

$$\min\{|\frac{\hat{\boldsymbol{w}}^T\boldsymbol{\mu}_1 + \hat{b}}{\sigma} - \rho/2|, |\frac{\hat{\boldsymbol{w}}^T\boldsymbol{\mu}_1 + \hat{b}}{\sigma} + \rho/2|\} \lesssim \frac{1}{\rho^2}\sqrt{\frac{p}{n}\log(pn)} \tag{64}$$

where we used the fact that $n \gtrsim p\log(pn)/\rho^6 \gtrsim p\log(pn)/\rho^4$.

Similarly, we simultaneously have

$$\min\{|\frac{\hat{\boldsymbol{w}}^T\boldsymbol{\mu}_1 + \hat{b}}{\sigma} - \rho/2|, |\frac{\hat{\boldsymbol{w}}^T\boldsymbol{\mu}_1 + \hat{b}}{\sigma} + \rho/2|\} \lesssim \frac{1}{\rho^2}\sqrt{\frac{p}{n}\log(pn)} \tag{65}$$

Because $n \gtrsim p\log(pn)/\rho^6 \gtrsim p$, we have $\rho \gtrsim \frac{1}{\rho^2}\sqrt{\frac{p}{n}\log(pn)}$.

Finally, using Corollary 2, when $n \gtrsim p\log(pn)/\rho^6 \gtrsim p$, we have

$$\mathbb{E}[\mathcal{E}(\hat{\mathsf{C}})] \lesssim \frac{1}{\rho^3}\frac{p}{n}\log(pn) \tag{66}$$

□

B.2 Sparse Setting

B.2.1 Upper Bound

Proof of Theorem 13 (Clustering upper bound). The proof is based on Proposition 9 of [1], which we state here for completeness.

Proposition 4 (Proposition 9, [1]). *Assume that $n \geq 1$, $p \geq 2$, and $\alpha < 1/4$. Define $\tilde{S} = \{i : |(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)_i| \geq 4\sigma\sqrt{\alpha}\}$. Then $\tilde{S} \subseteq \hat{S} \subseteq S$ with probability at least $1 - 6/n$.*

Notice that $\tilde{S} \neq \varnothing$ because $n \gtrsim \log(np)/\rho^4$.

Let

$$\begin{aligned}
\cos\hat{\beta} &= \boldsymbol{\mu}^T\hat{\boldsymbol{v}}/\|\boldsymbol{\mu}\|_2 && (67)\\
\cos\bar{\beta} &= \boldsymbol{\mu}^T\boldsymbol{\mu}_{\hat{S}}/\|\boldsymbol{\mu}\|_2\|\boldsymbol{\mu}_{\hat{S}}\|_2 && (68)\\
\cos\beta &= \boldsymbol{\mu}_{\hat{S}}^T\hat{\boldsymbol{v}}/\|\boldsymbol{\mu}_{\hat{S}}\|_2 && (69)
\end{aligned}$$

And we have

$$\begin{aligned}
\sin\bar{\beta} &= \frac{\|\boldsymbol{\mu}-\boldsymbol{\mu}_{\hat{S}}\|_2}{\|\boldsymbol{\mu}\|_2} && (70)\\
&\le \frac{\|\boldsymbol{\mu}-\boldsymbol{\mu}_{\tilde{S}}\|_2}{\|\boldsymbol{\mu}\|_2} && (71)\\
&\le \frac{4\sigma\sqrt{\alpha}\sqrt{s-|\tilde{S}|}}{\|\boldsymbol{\mu}\|_2} && (72)\\
&\lesssim \frac{\sqrt{s\alpha}}{\rho} && (73)
\end{aligned}$$

which is sufficiently small when $n \gtrsim s^2\log(np)/\rho^2$.

Using Proposition 6 of [1] (Proposition 1), with high probability, we have

$$\sin\beta \le \frac{1}{\rho^2}\sqrt{\frac{s\log(ns)}{n}} \qquad (74)$$

which is sufficiently small when $n \gtrsim s^2\log(np)/\rho^4$.

Thus, from the triangle inequality in spherical geometry, we have

$$\hat{\beta} \le \bar{\beta} + \beta \qquad (75)$$

which is sufficiently small. And we have

$$\begin{aligned}
\sin\hat{\beta} &\lesssim \sin\bar{\beta} + \sin\beta && (76)\\
&\lesssim \frac{1}{\rho^2}\sqrt{\frac{s\log(ns)}{n}} + \frac{\sqrt{s}}{\rho}(\frac{\log(pn)}{n})^{\frac{1}{4}} && (77)
\end{aligned}$$

Now, similar to the proof of Theorem 11, we have

$$\mathbb{E}[\mathcal{E}(\hat{\mathsf{C}})] \lesssim \frac{1}{\rho^3}\frac{s\log(ns)}{n} + \frac{s}{\rho^2}\left(\frac{\log(pn)}{n}\right)^{\frac{1}{2}} \qquad (78)$$

when $n \gtrsim s^2\log(pn)/\rho^8$.

□