
More Supervision, Less Computation: Statistical-Computational Tradeoffs in Weakly Supervised Learning

Xinyang Yi^{†*} Zhaoran Wang^{‡*} Zhuoran Yang^{‡*} Constantine Caramanis[†] Han Liu[‡]

[†]The University of Texas at Austin [‡]Princeton University

[†]{yixy, constantine}@utexas.edu [‡]{zhaoran, zy6, hanliu}@princeton.edu

{*: equal contribution}

Abstract

We consider the weakly supervised binary classification problem where the labels are randomly flipped with probability $1 - \alpha$. Although there exist numerous algorithms for this problem, it remains theoretically unexplored how the statistical accuracies and computational efficiency of these algorithms depend on the degree of supervision, which is quantified by α . In this paper, we characterize the effect of α by establishing the information-theoretic and computational boundaries, namely, the minimax-optimal statistical accuracy that can be achieved by all algorithms, and polynomial-time algorithms under an oracle computational model. For small α , our result shows a gap between these two boundaries, which represents the computational price of achieving the information-theoretic boundary due to the lack of supervision. Interestingly, we also show that this gap narrows as α increases. In other words, having more supervision, i.e., more correct labels, not only improves the optimal statistical accuracy as expected, but also enhances the computational efficiency for achieving such accuracy.

1 Introduction

Practical classification problems usually involve corrupted labels. Specifically, let $\{(\mathbf{x}_i, z_i)\}_{i=1}^n$ be n independent data points, where $\mathbf{x}_i \in \mathbb{R}^d$ is the covariate vector and $z_i \in \{0, 1\}$ is the uncorrupted label. Instead of observing $\{(\mathbf{x}_i, z_i)\}_{i=1}^n$, we observe $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ in which y_i is the corrupted label. In detail, with probability $(1 - \alpha)$, y_i is chosen uniformly at random over $\{0, 1\}$, and with probability α , $y_i = z_i$. Here $\alpha \in [0, 1]$ quantifies the degree of supervision: a larger α indicates more supervision since we have more uncorrupted labels in this case. In this paper, we are particularly interested in the effect of α on the statistical accuracy and computational efficiency for parameter estimation in this problem, particularly in the high dimensional settings where the dimension d is much larger than the sample size n .

There exists a vast body of literature on binary classification problems with corrupted labels. In particular, the study of randomly perturbed labels dates back to [1] in the context of random classification noise model. See, e.g., [12, 20] for a survey. Also, classification problems with missing labels are also extensively studied in the context of semi-supervised or weakly supervised learning by [14, 17, 21], among others. Despite the extensive study on this problem, its information-theoretic and computational boundaries remain unexplored in terms of theory. In a nutshell, the information-theoretic boundary refers to the optimal statistical accuracy achievable by any algorithms, while the computational boundary refers to the optimal statistical accuracy achievable by the algorithms under a computational budget that has a polynomial dependence on the problem scale (d, n) . Moreover, it remains unclear how these two boundaries vary along with α . One interesting question to ask is

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

how the degree of supervision affects the fundamental statistical and computational difficulties of this problem, especially in the high dimensional regime.

In this paper, we sharply characterize both the information-theoretic and computational boundaries of the weakly supervised binary classification problems under the minimax framework. Specifically, we consider the Gaussian generative model where $\mathbf{X}|Z = z \sim \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma})$ and $z \in \{0, 1\}$ is the true label. Suppose $\{(\mathbf{x}_i, z_i)\}_{i=1}^n$ are n independent samples of (\mathbf{X}, Z) . We assume that $\{y_i\}_{i=1}^n$ are generated from $\{z_i\}_{i=1}^n$ in the aforementioned manner. We focus on the high dimensional regime, where $d \gg n$ and $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ is s -sparse, i.e., $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ has s nonzero entries. We are interested in estimating $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ from the observed samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. By a standard reduction argument [24], the fundamental limits of this estimation task are captured by a hypothesis testing problem, namely, $H_0 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0 = \mathbf{0}$ versus $H_1 : \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$ is s -sparse and

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) := \gamma_n > 0, \quad (1.1)$$

where γ_n denotes the signal strength that scales with n . Consequently, we focus on studying the fundamental limits of γ_n for solving this hypothesis testing problem.

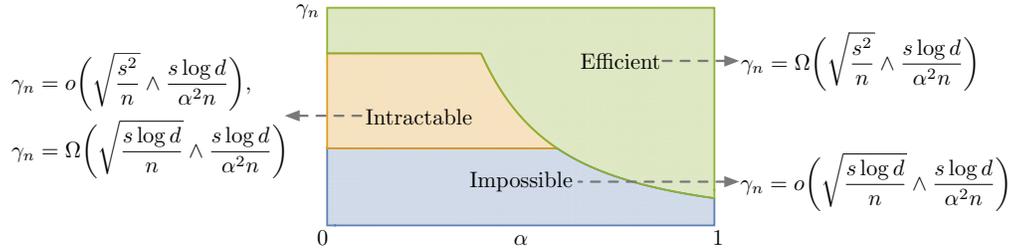


Figure 1: Computational-statistical phase transitions for weakly supervised binary classification. Here α denotes the degree of supervision, i.e., the label is corrupted to be uniformly random with probability $1 - \alpha$, and γ_n is the signal strength, which is defined in (1.1). Here $a \wedge b$ denotes $\min\{a, b\}$.

Our main results are illustrated in Figure 1. Specifically, we identify the impossible, intractable, and efficient regimes for the statistical-computational phase transitions under certain regularity conditions.

- (i) For $\gamma_n = o[\sqrt{s \log d/n} \wedge (1/\alpha^2 \cdot s \log d/n)]$, any algorithm is asymptotically powerless in solving the hypothesis testing problem.
- (ii) For $\gamma_n = \Omega[\sqrt{s \log d/n} \wedge (1/\alpha^2 \cdot s \log d/n)]$ and $\gamma_n = o[\sqrt{s^2/n} \wedge (1/\alpha^2 \cdot s \log d/n)]$, any tractable algorithm that has a polynomial oracle complexity under an extension of the statistical query model [18] is asymptotically powerless. We will rigorously define the computational model in §2.
- (iii) For $\gamma_n = \Omega[\sqrt{s^2/n} \wedge (1/\alpha^2 \cdot s \log d/n)]$, there is an efficient algorithm with a polynomial oracle complexity that is asymptotically powerful in solving the testing problem.

Here $\sqrt{s \log d/n} \wedge (1/\alpha^2 \cdot s \log d/n)$ gives the information-theoretic boundary, while $\sqrt{s^2/n} \wedge (1/\alpha^2 \cdot s \log d/n)$ gives the computational boundary. Moreover, by a reduction from the estimation problem to the testing problem, these boundaries for testing imply the ones for estimating $\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$ as well.

Consequently, there exists a significant gap between the computational and information-theoretic boundaries for small α . In other words, to achieve the information-theoretic boundary, one has to pay the price of intractable computation. As α tends to one, this gap between computational and information-theoretic boundaries narrows and eventually vanishes. This indicates that, having more supervision not only improves the statistical accuracy, as shown by the decay of information-theoretic boundary in Figure 1, but more importantly, enhances the computational efficiency by reducing the computational price for attaining information-theoretic optimality. This phenomenon — “more supervision, less computation” — is observed for the first time in this paper.

1.1 More Related Work, Our Contribution, and Notation

Besides the aforementioned literature on weakly supervised learning and label corruption, our work is also connected to a recent line of work on statistical-computational tradeoffs [2–5, 8, 13, 15, 19, 26–28]. In comparison, we quantify the statistical-computational tradeoffs for weakly supervised learning for the first time. Furthermore, our results are built on an oracle computational model

in [8] that slightly extends the statistical query model [18], and hence do not hinge on unproven conjectures on computational hardness like **planted clique**. Compared with our work, [8] focuses on the computational hardness of learning heterogeneous models, whereas we consider the interplay between supervision and statistical-computational tradeoffs. A similar computational model is used in [27] to study structural normal mean model and principal component analysis, which exhibit different statistical-computational phase transitions. In addition, our work is related to sparse linear discriminant analysis and two-sample testing of sparse means, which correspond to our special cases of $\alpha = 1$ and $\alpha = 0$, respectively. See, e.g., [7, 23] for details. In contrast with their results, our results capture the effects of α on statistical and computational tradeoffs.

In summary, the contribution of our work is two-fold:

- (i) We characterize the computational and statistical boundaries of the weakly supervised binary classification problem for the first time. Compared with existing results for other models, our results do not rely on unproven conjectures.
- (ii) Based on our theoretical characterization, we propose the “more supervision, less computation” phenomenon, which is observed for the first time.

Notation. We denote the χ^2 -divergence between two distributions \mathbb{P}, \mathbb{Q} by $D_{\chi^2}(\mathbb{P}, \mathbb{Q})$. For two nonnegative sequences a_n, b_n indexed by n , we use $a_n = o(b_n)$ as a shorthand for $\lim_{n \rightarrow \infty} a_n/b_n = 0$. We say $a_n = \Omega(b_n)$ if $a_n/b_n \geq c$ for some absolute constant $c > 0$ when n is sufficiently large. We use $a \vee b$ and $a \wedge b$ to denote $\max\{a, b\}$ and $\min\{a, b\}$, respectively. For any positive integer k , we denote $\{1, 2, \dots, k\}$ by $[k]$. For $\mathbf{v} \in \mathbb{R}^d$, we denote by $\|\mathbf{v}\|_p$ the ℓ_p -norm of \mathbf{v} . In addition, we denote the operator norm of a matrix \mathbf{A} by $\|\mathbf{A}\|_2$.

2 Background

In this section, we formally define the statistical model for weakly supervised binary classification. Then we follow it with the statistical query model that connects computational complexity and statistical optimality.

2.1 Problem Setup

Consider the following Gaussian generative model for binary classification. For a random vector $\mathbf{X} \in \mathbb{R}^d$ and a binary random variable $Z \in \{0, 1\}$, we assume

$$\mathbf{X}|Z=0 \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}), \quad \mathbf{X}|Z=1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \quad (2.1)$$

where $\mathbb{P}(Z=0) = \mathbb{P}(Z=1) = 1/2$. Under this model, the optimal classifier by Bayes rule corresponds to the Fisher’s linear discriminative analysis (LDA) classifier. In this paper, we focus on the noisy label setting where true label Z is replaced by a uniformly random label in $\{0, 1\}$ with probability $1 - \alpha$. Hence, α characterizes the degree of supervision in the model. In specific, if $\alpha = 0$, we observe the true label Z , thus the problem belongs to supervised learning. Whereas if $\alpha = 1$, the observed label is completely random, which contains no information of the model in (2.1). This setting is thus equivalent to learning a Gaussian mixture model, which is an unsupervised problem. In the general setting with noisy labels, we denote the observed label by Y , which is linked to the true label Z via

$$\mathbb{P}(Y=Z) = (1 + \alpha)/2, \quad \mathbb{P}(Y=1-Z) = (1 - \alpha)/2. \quad (2.2)$$

We consider the hypothesis testing problem of detecting whether $\boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1$ given n i.i.d. samples $\{y_i, \mathbf{x}_i\}_{i=1}^n$ of (Y, \mathbf{X}) , namely

$$H_0: \boldsymbol{\mu}_0 = \boldsymbol{\mu}_1 \text{ versus } H_1: \boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_1. \quad (2.3)$$

We focus on the high dimensional and sparse regime, where $d \gg n$ and $\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ is s -sparse, i.e., $\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1 \in \mathcal{B}_0(s)$, where $\mathcal{B}_0(s) := \{\boldsymbol{\mu} \in \mathbb{R}^d: \|\boldsymbol{\mu}\|_0 \leq s\}$. Throughout this paper, use the sample size n to drive the asymptotics. We introduce a shorthand notation $\boldsymbol{\theta} := (\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}, \alpha)$ to represent the parameters of the aforementioned model. Let $\mathbb{P}_{\boldsymbol{\theta}}$ be the joint distribution of (Y, \mathbf{X}) under our statistical model with parameter $\boldsymbol{\theta}$, and $\mathbb{P}_{\boldsymbol{\theta}}^n$ be the product distribution of n i.i.d. samples accordingly. We denote the parameter spaces of the null and alternative hypotheses by \mathcal{G}_0 and \mathcal{G}_1 respectively. For any test function $\phi: \{(y_i, \mathbf{x}_i)\}_{i=1}^n \rightarrow \{0, 1\}$, the classical testing risk is defined as the summation of

type-I and type-II errors, namely

$$R_n(\phi; \mathcal{G}_0, \mathcal{G}_1) := \sup_{\theta \in \mathcal{G}_0} \mathbb{P}_\theta^n(\phi = 1) + \sup_{\theta \in \mathcal{G}_1} \mathbb{P}_\theta^n(\phi = 0).$$

The minimax risk is defined as the smallest testing risk of all possible test functions, that is,

$$R_n^*(\mathcal{G}_0, \mathcal{G}_1) := \inf_{\phi} R_n(\phi; \mathcal{G}_0, \mathcal{G}_1), \quad (2.4)$$

where the infimum is taken over all measurable test functions.

Intuitively, the separation between two Gaussian components under H_1 and the covariance matrix Σ together determine the hardness of detection. To characterize such dependence, we define the signal-to-noise ratio (SNR) as $\rho(\theta) := (\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 - \mu_1)$. For any nonnegative sequence $\{\gamma_n\}_{n \geq 1}$, let $\mathcal{G}_1(\gamma_n) := \{\theta : \rho(\theta) \geq \gamma_n\}$ be a sequence of alternative parameter spaces with minimum separation γ_n . The following minimax rate characterizes the information-theoretic limits of the detection problem.

Definition 2.1 (Minimax rate). *We say a sequence $\{\gamma_n^*\}_{n \geq 1}$ is a minimax rate if*

- For any sequence $\{\gamma_n\}_{n \geq 1}$ satisfying $\gamma_n = o(\gamma_n^*)$, we have $\lim_{n \rightarrow \infty} R_n^*[\mathcal{G}_0, \mathcal{G}_1(\gamma_n)] = 1$;
- For any sequence $\{\gamma_n\}_{n \geq 1}$ satisfying $\gamma_n = \Omega(\gamma_n^*)$, we have $\lim_{n \rightarrow \infty} R_n^*[\mathcal{G}_0, \mathcal{G}_1(\gamma_n)] = 0$.

The minimax rate in Definition 2.1 characterizes the statistical difficulty of the testing problem. However, it fails to shed light on the computational efficiency of possible testing algorithms. The reason is that this concept does not make any computational restriction on the test functions. The minimax risk in (2.4) might be attained only by test functions that have exponential computational complexities. This limitation of Definition 2.1 motivates us to study statistical limits under computational constraints.

2.2 Computational Model

Statistical query models [8–11, 18, 27] capture computational complexity by characterizing the total number of rounds an algorithm interacts with data. In this paper, we consider the following statistical query model, which admits bounded query functions but allows the responses of query functions to be unbounded.

Definition 2.2 (Statistical query model). *In the statistical query model, an algorithm \mathcal{A} is allowed to query an oracle T rounds, but not to access data $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$ directly. At each round, \mathcal{A} queries the oracle r with a query function $q \in \mathcal{Q}_{\mathcal{A}}$, in which $\mathcal{Q}_{\mathcal{A}} \subseteq \{q : \{0, 1\} \times \mathbb{R}^d \rightarrow [-M, M]\}$ denotes the query space of \mathcal{A} . The oracle r outputs a realization of a random variable $Z_q \in \mathbb{R}$ satisfying*

$$\mathbb{P}\left(\bigcap_{q \in \mathcal{Q}_{\mathcal{A}}} \{|Z_q - \mathbb{E}[q(Y, \mathbf{X})]|\} \leq \tau_q\right) \geq 1 - 2\xi, \quad \text{where}$$

$$\tau_q = [\eta(\mathcal{Q}_{\mathcal{A}}) + \log(1/\xi)] \cdot M/n \sqrt{2[\eta(\mathcal{Q}_{\mathcal{A}}) + \log(1/\xi)] \cdot (M^2 - \{\mathbb{E}[q(Y, \mathbf{X})]\}^2)/n}. \quad (2.5)$$

Here $\tau_q > 0$ is the tolerance parameter and $\xi \in [0, 1)$ is the tail probability. The quantity $\eta(\mathcal{Q}_{\mathcal{A}}) \geq 0$ in τ_q measures the capacity of $\mathcal{Q}_{\mathcal{A}}$ in logarithmic scale, e.g., for countable $\mathcal{Q}_{\mathcal{A}}$, $\eta(\mathcal{Q}_{\mathcal{A}}) = \log(|\mathcal{Q}_{\mathcal{A}}|)$. The number T is defined as the oracle complexity. We denote by $\mathcal{R}[\xi, n, T, \eta(\mathcal{Q}_{\mathcal{A}})]$ the set of oracles satisfying (2.5), and by $\mathcal{A}(T)$ the family of algorithms that queries an oracle no more than T rounds.

This version of statistical query model is used in [8], and reduces to the VSTAT model proposed in [9–11] by the transformation $\tilde{q}(y, \mathbf{x}) = q(y, \mathbf{x})/(2M) + 1/2$ for any $q \in \mathcal{Q}_{\mathcal{A}}$. The computational model in Definition 2.2 enables us to handle query functions that are bounded by an unknown and fixed number M . Note that that by incorporating the tail probability ξ , the response Z_q is allowed to be unbounded. To understand the intuition behind Definition 2.2, we remark that (2.5) resembles the Bernstein’s inequality for bounded random variables [25]

$$\mathbb{P}\left\{\left|\frac{1}{n} \sum_{i=1}^n q(Y_i, \mathbf{X}_i) - \mathbb{E}[q(Y, \mathbf{X})]\right| \geq t\right\} \leq 2 \exp\left\{-\frac{t^2}{2\text{Var}[q(Y, \mathbf{X})] + Mt}\right\}. \quad (2.6)$$

We first replace $\text{Var}[q(Y, \mathbf{X})]$ by its upper bound $M^2 - \{\mathbb{E}[q(Y, \mathbf{X})]\}^2$, which is tight when q takes values in $\{-M, M\}$. Then inequality (2.5) is obtained by replacing $n^{-1} \sum_{i=1}^n q(Y_i, \mathbf{X}_i)$ in (2.6) by Z_q and then bounding the suprema over the query space $\mathcal{Q}_{\mathcal{A}}$. In the definition of τ_q in (2.5), we

incorporate the effect of uniform concentration over the query space $\mathcal{Q}_{\mathcal{A}}$ by adding the quantity $\eta(\mathcal{Q}_{\mathcal{A}})$, which measures the capacity of $\mathcal{Q}_{\mathcal{A}}$. In addition, under the Definition 2.2, the algorithm \mathcal{A} does not interact directly with data. Such a restriction characterizes the fact that in statistical problems, the effectiveness of an algorithm only depends on the global statistical properties, not the information of individual data points. For instance, algorithms that only rely on the convergence of the empirical distribution to the population distribution are contained in the statistical query model; whereas algorithms that hinge on the first data point (y_1, \mathbf{x}_1) is not allowed. This restriction captures a vast family of algorithms in statistics and machine learning, including applying gradient method to maximize likelihood function, matrix factorization algorithms, expectation-maximization algorithms, and sampling algorithms [9].

Based on the statistical query model, we study the minimax risk under oracle complexity constraints. For the testing problem (2.3), let $\mathcal{A}(T_n)$ be a class of testing algorithms under the statistical query model with query complexity no more than T_n , with $\{T_n\}_{n \geq 1}$ being a sequence of positive integers depending on the sample size n . For any $\mathcal{A} \in \mathcal{A}(T_n)$ and any oracle $r \in \mathcal{R}[\xi, n, T_n, \eta(\mathcal{Q}_{\mathcal{A}})]$ that responds to \mathcal{A} , let $\mathcal{H}(\mathcal{A}, r)$ be the set of test functions that deterministically depend on \mathcal{A} 's queries to the oracle r and the corresponding responses. We use $\bar{\mathbb{P}}_{\theta}$ to denote the distribution of the random variables returned by oracle r when the model parameter is θ .

For a general hypothesis testing problem, namely, $H_0: \theta \in \mathcal{G}_0$ versus $H_1: \theta \in \mathcal{G}_1$, the minimax testing risk with respect to an algorithm \mathcal{A} and a statistical oracle $r \in \mathcal{R}[\xi, n, T_n, \eta(\mathcal{Q}_{\mathcal{A}})]$ is defined as

$$\bar{R}_n^*(\mathcal{G}_0, \mathcal{G}_1; \mathcal{A}, r) := \inf_{\phi \in \mathcal{H}(\mathcal{A}, r)} \left[\sup_{\theta \in \mathcal{G}_0} \bar{\mathbb{P}}_{\theta}(\phi = 1) + \sup_{\theta \in \mathcal{G}_1} \bar{\mathbb{P}}_{\theta}(\phi = 0) \right]. \quad (2.7)$$

Compared with the classical minimax risk in (2.4), the new notion in (2.7) incorporates the computational budgets via oracle complexity. In specific, we only consider the test functions obtained by an algorithm with at most T_n queries to a statistical oracle. If T_n is a polynomial of the dimensionality d , (2.7) characterizes the statistical optimality of computational efficient algorithms. This motivates us to define the computationally tractable minimax rate, which contrasts with Definition 2.1.

Definition 2.3 (Computationally tractable minimax rate). *Let $\mathcal{G}_1(\gamma_n) := \{\theta : \rho(\theta) \geq \gamma_n\}$ be a sequence of model spaces with minimum separation γ_n , where $\rho(\theta)$ is the SNR. A sequence $\{\bar{\gamma}_n\}_{n \geq 1}$ is called a computationally tractable minimax rate if*

- For any sequence $\{\gamma_n\}_{n \geq 1}$ satisfying $\gamma_n = o(\bar{\gamma}_n^*)$, any constant $\eta > 0$, and any $\mathcal{A} \in \mathcal{A}(d^n)$, there exists an oracle $r \in \mathcal{R}[\xi, n, T_n, \eta(\mathcal{Q}_{\mathcal{A}})]$ such that $\lim_{n \rightarrow \infty} \bar{R}_n^*[\mathcal{G}_0, \mathcal{G}_1(\gamma_n); \mathcal{A}, r] = 1$;
- For any sequence $\{\gamma_n\}_{n \geq 1}$ satisfying $\gamma_n = \Omega(\bar{\gamma}_n^*)$, there exist a constant $\eta > 0$ and an algorithm $\mathcal{A} \in \mathcal{A}(d^n)$ such that, for any $r \in \mathcal{R}[\xi, n, T_n, \eta(\mathcal{Q}_{\mathcal{A}})]$, we have $\lim_{n \rightarrow \infty} \bar{R}_n^*[\mathcal{G}_0, \mathcal{G}_1(\gamma_n); \mathcal{A}, r] = 0$.

3 Main Results

Throughout this paper, we assume that the covariance matrix Σ in (2.1) is known. Specifically, for some positive definite $\Sigma \in \mathbb{R}^{d \times d}$, the parameter spaces of the null and alternative hypotheses are defined as

$$\mathcal{G}_0(\Sigma) := \{\theta = (\mu, \mu, \Sigma, \alpha) : \mu \in \mathbb{R}^d\}, \quad (3.1)$$

$$\mathcal{G}_1(\Sigma; \gamma_n) := \{\theta = (\mu_0, \mu_1, \Sigma, \alpha) : \mu_0, \mu_1 \in \mathbb{R}^d, \mu_0 - \mu_1 \in \mathcal{B}_0(s), \rho(\theta) \geq \gamma_n\}. \quad (3.2)$$

Accordingly, the testing problem of detecting whether $\mu_0 \neq \mu_1$ is to distinguish

$$H_0: \theta \in \mathcal{G}_0(\Sigma) \text{ versus } H_1: \theta \in \mathcal{G}_1(\Sigma; \gamma_n). \quad (3.3)$$

In §3.1, we present the minimax rate of the detection problem from an information-theoretic perspective. In §3.2, under the statistical query model introduced in §2.2, we provide a computational lower bound and a nearly matching upper bound that is achieved by an efficient testing algorithm.

3.1 Information-theoretic Limits

Now we turn to characterize the minimax rate given in Definition 2.1. For parameter spaces (3.1) and (3.2) with known Σ , we show that in highly sparse setting where $s = o(\sqrt{d})$, we have

$$\gamma_n^* = \sqrt{s \log d/n} \wedge (1/\alpha^2 \cdot s \log d/n), \quad (3.4)$$

To prove (3.4), we first present a lower bound which shows that the hypothesis testing problem in (3.3) is impossible if $\gamma_n = o(\gamma_n^*)$.

Theorem 3.1. *For the hypothesis testing problem in (3.3) with known Σ , we assume that there exists a small constant $\delta > 0$ such that $s = o(d^{1/2-\delta})$. Let γ_n^* be defined in (3.4). For any sequence $\{\gamma_n\}_{n \geq 1}$ such that $\gamma_n = o(\gamma_n^*)$, any hypothesis test is asymptotically powerless, namely,*

$$\lim_{n \rightarrow \infty} \sup_{\Sigma} R_n^*[\mathcal{G}_0(\Sigma), \mathcal{G}_1(\Sigma; \gamma_n)] = 1.$$

By Theorem 3.1, we observe a phase transition in the necessary SNR for powerful detection when α decreases from one to zero. Starting with rate $s \log d/n$ in the supervised setting where $\alpha = 1$, the required SNR gradually increases as label qualities decrease. Finally, when α reaches zero, which corresponds to the unsupervised setting, powerful detection requires the SNR to be $\Omega(\sqrt{s \log d/n})$. It is worth noting that when $\alpha = (s \log d/n)^{1/4}$, we still have $(n^3 s \log d)^{1/4}$ uncorrupted labels. However, our lower bound (along with the upper bound shown in Theorem 3.2) indicates that the information contained in these uncorrupted labels are buried in the noise, and cannot essentially improve the detection quality compared with the unsupervised setting.

Next we establish a matching upper bound for the detection problem in (3.3). We denote the condition number of the covariance matrix Σ by κ , i.e., $\kappa := \lambda_{\max}(\Sigma)/\lambda_{\min}(\Sigma)$, where $\lambda_{\max}(\Sigma)$ and $\lambda_{\min}(\Sigma)$ are the largest and smallest eigenvalues of Σ , respectively. Note that marginally Y is uniformly distributed over $\{0, 1\}$. For ease of presentation, we assume that the sample size is $2n$ and each class contains exactly n data points. Note that we can always discard some samples in the larger class to make the sample sizes of both classes to be equal. Due to the law of large numbers, this trick will not affect the analysis of sample complexity in the sense of order wise.

Given $2n$ i.i.d. samples $\{(y_i, \mathbf{x}_i)\}_{i=1}^{2n}$ of $(Y, \mathbf{X}) \in \{0, 1\} \times \mathbb{R}^d$, we define

$$\mathbf{w}_i = \Sigma^{-1/2}(\mathbf{x}_{2i} - \mathbf{x}_{2i-1}), \text{ for all } i \in [n]. \quad (3.5)$$

In addition, we split the dataset $\{(y_i, \mathbf{x}_i)\}_{i=1}^{2n}$ into two disjoint parts $\{(0, \mathbf{x}_i^{(0)})\}_{i=1}^n$ and $\{(1, \mathbf{x}_i^{(1)})\}_{i=1}^n$, and define

$$\mathbf{u}_i = \mathbf{x}_i^{(1)} - \mathbf{x}_i^{(0)}, \text{ for all } i \in [n]. \quad (3.6)$$

We note that computing sample differences in (3.5) and (3.6) is critical for our problem because we focus on detecting the difference between μ_0 and μ_1 , and computing differences can avoid estimating $\mathbb{E}_{\mathbb{P}_\theta}(\mathbf{X})$ that might be dense. For any integer $s \in [d]$, we define $\mathcal{B}_2(s) := \mathcal{B}_0(s) \cap \mathbb{S}^{d-1}$ as the set of s -sparse vectors on the unit sphere in \mathbb{R}^d . With $\{\mathbf{w}_i\}_{i=1}^n$ and $\{\mathbf{u}_i\}_{i=1}^n$, we introduce two test functions

$$\phi_1 := \mathbb{1} \left\{ \sup_{\mathbf{v} \in \mathcal{B}_2(s)} \frac{1}{n} \sum_{i=1}^n \frac{(\mathbf{v}^\top \Sigma^{-1} \mathbf{w}_i)^2}{2\mathbf{v}^\top \Sigma^{-1} \mathbf{v}} \geq 1 + \tau_1 \right\}, \quad (3.7)$$

$$\phi_2 := \mathbb{1} \left\{ \sup_{\mathbf{v} \in \mathcal{B}_2(1)} \frac{1}{n} \sum_{i=1}^n \langle \mathbf{v}, \text{diag}(\Sigma)^{-1/2} \mathbf{u}_i \rangle \geq \tau_2 \right\}, \quad (3.8)$$

where $\tau_1, \tau_2 > 0$ are algorithmic parameters that will be specified later. To provide some intuitions, we consider the case where $\Sigma = \mathbf{I}$. Test function ϕ_1 seeks a sparse direction that explains the most variance of \mathbf{w}_i . Therefore, such a test is closely related to the sparse principal component detection problem [3]. Test function ϕ_2 simply selects the coordinate of $n^{-1} \sum_{i=1}^n \mathbf{u}_i$ that has the largest magnitude and compares it with τ_2 . This test is closely related to detecting sparse normal mean in high dimensions [16]. Based on these two ingredients, we construct our final testing function ϕ as $\phi = \phi_1 \vee \phi_2$, i.e., if any of ϕ_1 and ϕ_2 is true, then ϕ rejects the null. The following theorem establishes a sufficient condition for test function ϕ to be asymptotically powerful.

Theorem 3.2. *Consider the testing problem (3.3) where Σ is known and has condition number κ . For test functions ϕ_1 and ϕ_2 defined in (3.7) and (3.8) with parameters τ_1 and τ_2 given by*

$$\tau_1 = \kappa \sqrt{s \log(ed/s)/n}, \quad \tau_2 = \sqrt{8 \log d/n}.$$

We define the ultimate test function as $\phi = \phi_1 \vee \phi_2$. We assume that $s \leq C \cdot d$ for some absolute constant C and $n \geq 64 \cdot s \log(ed/s)$. Then if

$$\gamma_n \geq C' \kappa \cdot [\sqrt{s \log(ed/s)/n} \wedge (1/\alpha^2 \cdot s \log d/n)], \quad (3.9)$$

where C' is an absolute constant, then test function ϕ is asymptotically powerful. In specific, we have

$$\sup_{\theta \in \mathcal{G}_0(\Sigma)} \mathbb{P}_\theta^n(\phi = 1) + \sup_{\theta \in \mathcal{G}_1(\Sigma; \gamma_n)} \mathbb{P}_\theta^n(\phi = 0) \leq 20/d. \quad (3.10)$$

Theorem 3.2 provides a non-asymptotic guarantee. When n goes to infinity, (3.10) implies that the test function ϕ is asymptotically powerful. When $s = o(\sqrt{d})$ and κ is a constant, (3.9) yields $\gamma_n = \Omega[\sqrt{s \log d/n} \wedge (1/\alpha^2 \cdot s \log d/n)]$, which matches the lower bound given in Theorem 3.1. Thus we conclude that γ_n^* defined in (3.4) is the minimax rate of testing problem in (3.3). We remark that when $s = \Omega(d)$, $\alpha = 1$, i.e., the standard (low-dimensional) setting of two sample testing, the bound provided in (3.9) is sub-optimal as [22] shows that SNR rate \sqrt{d}/n is sufficient for asymptotically powerful detection when $n = \Omega(\sqrt{d})$. It is thus worth noting that we focus on the highly sparse setting $s = o(\sqrt{d})$ and provided sharp minimax rate for this regime. In the definition of ϕ_1 in (3.7), we search over the set $\mathcal{B}_2(s)$. Since $\mathcal{B}_2(s)$ contains $\binom{d}{s}$ distinct sets of supports, computing ϕ_1 requires exponential running time.

3.2 Computational Limits

In this section, we characterize the computationally tractable minimax rate $\bar{\gamma}_n^*$ given in Definition 2.3. Moreover, we focus on the setting where Σ is known a priori and the parameter spaces for the null and alternative hypotheses are defined in (3.1) and (3.2), respectively. The main result is that, in highly sparse setting where $s = o(\sqrt{d})$, we have

$$\bar{\gamma}_n^* = \sqrt{s^2/n} \wedge (1/\alpha^2 \cdot s \log d/n). \quad (3.11)$$

We first present the lower bound in the next result.

Theorem 3.3. *For the testing problem in (3.3) with Σ known a priori, we make the same assumptions as in Theorem 3.1. For any sequence $\{\gamma_n\}_{n \geq 1}$ such that*

$$\gamma_n = o \left\{ \gamma_n^* \vee \left[\sqrt{s^2/n} \wedge (1/\alpha^2 \cdot s/n) \right] \right\}, \quad (3.12)$$

where γ_n^* is defined in (3.4), any computationally tractable test is asymptotically powerless under the statistical query model. That is, for any constant $\eta > 0$ and any $\mathcal{A} \in \mathcal{A}(d^n)$, there exists an oracle $r \in \mathcal{R}[\xi, n, T_n, \eta(\mathcal{Q}_{\mathcal{A}})]$ such that $\lim_{n \rightarrow \infty} \bar{R}_n^*[\mathcal{G}_0(\Sigma), \mathcal{G}_1(\Sigma, \gamma_n); \mathcal{A}, r] = 1$.

We remark that the lower bound in (3.12) differs from γ_n^* in (3.11) by a logarithmic term when $\sqrt{1/n} \leq \alpha^2 \leq \sqrt{s \log d/n}$. We expect this gap to be eliminated by more delicate analysis under the statistical query model.

Now putting Theorems 3.1 and 3.3 together, we describe the ‘‘more supervision, less computation’’ phenomenon as follows.

- (i) When $0 \leq \alpha \leq (\log^2 d/n)^{1/4}$, the computational lower bound implies that the uncorrupted labels are unable to improve the quality of computationally tractable detection compared with the unsupervised setting. In addition, in this region, the gap between γ_n^* and $\bar{\gamma}_n^*$ remains the same.
- (ii) When $(\log^2 d/n)^{1/4} < \alpha \leq (s \log d/n)^{1/4}$, the information-theoretic lower bound shows that the uncorrupted labels cannot improve the quality of detection compared with unsupervised setting. However, more uncorrupted labels improve the statistical performances of hypothesis tests that are computationally tractable by shrinking the gap between γ_n^* and $\bar{\gamma}_n^*$.
- (iii) When $(s \log d/n)^{1/4} < \alpha \leq 1$, having more uncorrupted labels improves both statistical optimality and the computational efficiency. In specific, in this case, the gap between γ_n^* and $\bar{\gamma}_n^*$ vanishes and we have $\gamma_n^* = \bar{\gamma}_n^* = 1/\alpha^2 \cdot s \log d/n$.

Now we derive a nearly matching upper bound under the statistical query model, which establishes the computationally tractable minimax rate together with Theorem 3.3. We construct a computationally efficient testing procedure that combines two test functions which yields the two parts in $\bar{\gamma}_n^*$ respectively. Similar to ϕ_1 defined in (3.7), the first test function discards the information of labels, which works for the purely unsupervised setting where $\alpha = 0$. For $j \in [d]$, we denote by σ_j the j -th diagonal element of Σ . Under the statistical query model, we consider the $2d$ query functions

$$q_j(y, \mathbf{x}) := x_j / \sqrt{\sigma_j} \cdot \mathbb{1}\{|x_j / \sqrt{\sigma_j}| \leq R \cdot \sqrt{\log d}\}, \quad (3.13)$$

$$\tilde{q}_j(y, \mathbf{x}) := (x_j^2 / \sigma_j - 1) \cdot \mathbb{1}\{|x_j / \sqrt{\sigma_j}| \leq R \cdot \sqrt{\log d}\}, \text{ for all } j \in [d], \quad (3.14)$$

where $R > 0$ is an absolute constant. Here we apply truncation to the query functions to obtain bounded queries, which is specified by the statistical query model in Definition 2.2. We denote by z_{q_j} and $z_{\tilde{q}_j}$ the realizations of the random variables output by the statistical oracle for query functions q_j and \tilde{q}_j , respectively. As for the second test function, similar to (3.8), we consider

$$\bar{q}_{\mathbf{v}}(y, \mathbf{x}) = (2y - 1) \cdot \mathbf{v}^\top \text{diag}(\boldsymbol{\Sigma})^{-1/2} \mathbf{x} \cdot \mathbb{1}\{|\mathbf{v}^\top \text{diag}(\boldsymbol{\Sigma})^{-1/2} \mathbf{x}| \leq R \cdot \sqrt{\log d}\} \quad (3.15)$$

for all $\mathbf{v} \in \mathcal{B}_2(1)$. We denote by $Z_{\bar{q}_{\mathbf{v}}}$ the output of the statistical oracle corresponding to query function $\bar{q}_{\mathbf{v}}$. With these $4d$ query functions, we introduce test functions

$$\bar{\phi}_1 := \mathbb{1}\left\{\sup_{j \in [d]} (z_{\tilde{q}_j} - z_{q_j}^2) \geq C\bar{\tau}_1\right\}, \quad \bar{\phi}_2 := \mathbb{1}\left\{\sup_{\mathbf{v} \in \mathcal{B}_2(1)} z_{\bar{q}_{\mathbf{v}}} \geq 2\bar{\tau}_2\right\}, \quad (3.16)$$

where $\bar{\tau}_1$ and $\bar{\tau}_2$ are positive parameters that will be specified later and C is an absolute constant.

Theorem 3.4. *For the test functions $\bar{\phi}_1$ and $\bar{\phi}_2$ defined in (3.16), we define the ultimate test function as $\bar{\phi} = \bar{\phi}_1 \vee \bar{\phi}_2$. We set*

$$\bar{\tau}_1 = R^2 \log d \cdot \sqrt{\log(4d/\xi)/n}, \quad \bar{\tau}_2 = R \sqrt{\log d} \cdot \sqrt{\log(4d/\xi)/n}, \quad (3.17)$$

where $\xi = o(1)$. For the hypothesis testing problem in (3.3), we further assume that $\|\boldsymbol{\mu}_0\|_\infty \vee \|\boldsymbol{\mu}_1\|_\infty \leq C_0$ for some constant $C_0 > 0$. Under the assumption that

$$\sup_{j \in [d]} (\mu_{0,j} - \mu_{1,j})^2 / \sigma_j = \Omega\left\{[1/\alpha^2 \cdot \log^2 d \cdot \log(d/\xi)/n] \wedge \log d \cdot \sqrt{\log(d/\xi)/n}\right\}, \quad (3.18)$$

the risk of $\bar{\phi}$ satisfies that $\bar{R}_n^*(\bar{\phi}) = \sup_{\boldsymbol{\theta} \in \mathcal{G}_0(\boldsymbol{\Sigma})} \bar{\mathbb{P}}_{\boldsymbol{\theta}}(\bar{\phi} = 1) + \sup_{\boldsymbol{\theta} \in \mathcal{G}_1(\boldsymbol{\Sigma}, \gamma_n)} \bar{\mathbb{P}}_{\boldsymbol{\theta}}(\bar{\phi} = 0) \leq 5\xi$. Here we denote by $\mu_{0,j}$ and $\mu_{1,j}$ the j -th entry of $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$, respectively.

If we set the tail probability of the statistical query model to be $\xi = 1/d$, (3.18) shows that $\bar{\phi}$ is asymptotically powerful if $\sup_{j \in [d]} (\mu_{0,j} - \mu_{1,j})^2 / \sigma_j = \Omega[(1/\alpha^2 \cdot \log^3 d/n) \wedge (\log^3 d/n)^{1/2}]$.

When the energy of $\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ is spread over its support, $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_\infty$ and $\|\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1\|_2 / \sqrt{s}$ are close. Under the assumption that the condition number κ of $\boldsymbol{\Sigma}$ is a constant, (3.18) is implied by

$$\gamma_n \gtrsim (s^2 \log^3 d/n)^{1/2} \wedge (1/\alpha^2 \cdot s \log^3 d/n).$$

Compared with Theorem 3.3, the above upper bound matches the computational lower bound up to a logarithmic factor and $\bar{\gamma}_n^*$ is between $\sqrt{s^2/n} \wedge (1/\alpha^2 \cdot s \log d/n)$ and $(s^2 \log^3 d/n)^{1/2} \wedge (1/\alpha^2 \cdot s \log^3 d/n)$. Note that the truncation on query functions in (3.13) and (3.14) yields an additional logarithmic term, which could be reduced to $(s^2 \log d/n)^{1/2} \wedge (1/\alpha^2 \cdot s \log d/n)$ using more delicate analysis. Moreover, the test function $\bar{\phi}_1$ is essentially based on a diagonal thresholding algorithm performed on the covariance matrix of \mathbf{X} . The work in [6] provides a more delicate analysis of this algorithm which establishes the $\sqrt{s^2/n}$ rate. Their algorithm can also be formulated into the statistical query model; we use the simpler version in (3.16) for ease of presentation. Therefore, with more sophisticated proof technique, it can be shown that $\sqrt{s^2/n} \wedge (1/\alpha^2 \cdot s \log d/n)$ is the critical threshold for asymptotically powerful detection with computational efficiency.

3.3 Implication for Estimation

Our aforementioned phase transition in the detection problems directly implies the statistical and computational trade-offs in the problem of estimation. We consider the problem of estimating the parameter $\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_0 - \boldsymbol{\mu}_1$ of the binary classification model in (2.1) and (2.2), where $\Delta\boldsymbol{\mu}$ is s -sparse and $\boldsymbol{\Sigma}$ is known a priori. We assume that the signal to noise ratio is $\rho(\boldsymbol{\theta}) = \Delta\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \Delta\boldsymbol{\mu} \geq \gamma_n = o(\bar{\gamma}_n^*)$. For any constant $\eta > 0$ and any $\mathcal{A} \in \mathcal{A}(T)$ with $T = O(d^\eta)$, suppose we obtain an estimator $\Delta\hat{\boldsymbol{\mu}}$ of $\Delta\boldsymbol{\mu}$ by algorithm \mathcal{A} under the statistical query model. If $\Delta\hat{\boldsymbol{\mu}}$ converges to $\Delta\boldsymbol{\mu}$ in the sense that

$$(\Delta\hat{\boldsymbol{\mu}} - \Delta\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\Delta\hat{\boldsymbol{\mu}} - \Delta\boldsymbol{\mu}) = o[\gamma_n^2 / \rho(\boldsymbol{\theta})],$$

we have $|\Delta\hat{\boldsymbol{\mu}}^\top \boldsymbol{\Sigma}^{-1} \Delta\hat{\boldsymbol{\mu}} - \Delta\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \Delta\boldsymbol{\mu}| = o(\gamma_n)$. Thus the test function $\phi = \mathbb{1}\{\Delta\hat{\boldsymbol{\mu}}^\top \boldsymbol{\Sigma} \Delta\hat{\boldsymbol{\mu}} \geq \gamma_n/2\}$ is asymptotically powerful, which contradicts the computational lower bound in Theorem 3.3. Therefore, there exists a constant C such that $(\Delta\hat{\boldsymbol{\mu}} - \Delta\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\Delta\hat{\boldsymbol{\mu}} - \Delta\boldsymbol{\mu}) \geq C\gamma_n^2 / \rho(\boldsymbol{\theta})$ for any estimator $\Delta\hat{\boldsymbol{\mu}}$ constructed from polynomial number of queries.

Acknowledgments

We would like to thank Vitaly Feldman for valuable discussions.

References

- [1] Angluin, D. and Laird, P. (1988). Learning from noisy examples. *Machine Learning*, **2** 343–370.
- [2] Berthet, Q. and Rigollet, P. (2013). Computational lower bounds for sparse PCA. In *Conference on Learning Theory*.
- [3] Berthet, Q. and Rigollet, P. (2013). Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, **41** 1780–1815.
- [4] Chandrasekaran, V. and Jordan, M. I. (2013). Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, **110** 1181–1190.
- [5] Chen, Y. and Xu, J. (2014). Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *arXiv preprint arXiv:1402.1267*.
- [6] Deshpande, Y. and Montanari, A. (2014). Sparse PCA via covariance thresholding. In *Advances in Neural Information Processing Systems*.
- [7] Fan, J., Feng, Y. and Tong, X. (2012). A road to classification in high dimensional space: The regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B*, **74** 745–771.
- [8] Fan, J., Liu, H., Wang, Z. and Yang, Z. (2016). Curse of heterogeneity: Computational barriers in sparse mixture models and phase retrieval. *Manuscript*.
- [9] Feldman, V., Grigorescu, E., Reyzin, L., Vempala, S. and Xiao, Y. (2013). Statistical algorithms and a lower bound for detecting planted cliques. In *ACM Symposium on Theory of Computing*.
- [10] Feldman, V., Guzman, C. and Vempala, S. (2015). Statistical query algorithms for stochastic convex optimization. *arXiv preprint arXiv:1512.09170*.
- [11] Feldman, V., Perkins, W. and Vempala, S. (2015). On the complexity of random satisfiability problems with planted solutions. In *ACM Symposium on Theory of Computing*.
- [12] Fréney, B. and Verleysen, M. (2014). Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, **25** 845–869.
- [13] Gao, C., Ma, Z. and Zhou, H. H. (2014). Sparse CCA: Adaptive estimation and computational barriers. *arXiv preprint arXiv:1409.8565*.
- [14] Garcia-Garcia, D. and Williamson, R. C. (2011). Degrees of supervision. In *Advances in Neural Information Processing Systems*.
- [15] Hajek, B., Wu, Y. and Xu, J. (2014). Computational lower bounds for community detection on random graphs. *arXiv preprint arXiv:1406.6625*.
- [16] Johnstone, I. M. (1994). On minimax estimation of a sparse normal mean vector. *The Annals of Statistics*, **22** 271–289.
- [17] Joulin, A. and Bach, F. R. (2012). A convex relaxation for weakly supervised classifiers. In *International Conference on Machine Learning*.
- [18] Kearns, M. (1993). Efficient noise-tolerant learning from statistical queries. In *ACM Symposium on Theory of Computing*.
- [19] Ma, Z. and Wu, Y. (2014). Computational barriers in minimax submatrix detection. *The Annals of Statistics*, **43** 1089–1116.
- [20] Nettleton, D. F., Orriols-Puig, A. and Fornells, A. (2010). A study of the effect of different types of noise on the precision of supervised learning techniques. *Artificial Intelligence Review*, **33** 275–306.
- [21] Patrini, G., Nielsen, F., Nock, R. and Carioni, M. (2016). Loss factorization, weakly supervised learning and label noise robustness. *arXiv preprint arXiv:1602.02450*.
- [22] Ramdas, A., Singh, A. and Wasserman, L. (2016). Classification accuracy as a proxy for two sample testing. *arXiv preprint arXiv:1602.02210*.
- [23] Tony Cai, T., Liu, W. and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B*, **76** 349–372.
- [24] Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Springer.
- [25] Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.
- [26] Wang, T., Berthet, Q. and Samworth, R. J. (2014). Statistical and computational trade-offs in estimation of sparse principal components. *arXiv preprint arXiv:1408.5369*.
- [27] Wang, Z., Gu, Q. and Liu, H. (2015). Sharp computational-statistical phase transitions via oracle computational model. *arXiv preprint arXiv:1512.08861*.
- [28] Zhang, Y., Wainwright, M. J. and Jordan, M. I. (2014). Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*.

A Proofs of the Main Results

A.1 Proof of Theorem 3.1

In this section, we prove the information-theoretic lower bound. In specific, we focus on the restricted testing problem

$$H_0 : \boldsymbol{\theta} = (\mathbf{0}, \mathbf{0}, \mathbf{I}, \alpha) \text{ versus. } H_1 : \boldsymbol{\theta} = (-\mathbf{v}/2, \mathbf{v}/2, \mathbf{I}, \alpha), \quad (\text{A.1})$$

where

$$\mathbf{v} \in \mathcal{H}(s) := \{\mathbf{u} \in \{0, \beta\}^d : \|\mathbf{u}\|_0 = s\}.$$

Here we set $s\beta^2 = \gamma_n$ to ensure that $(-\mathbf{v}/2, \mathbf{v}/2, \mathbf{I}, \alpha)$ belongs to the alternative parameter space $\mathcal{G}(\boldsymbol{\Sigma}; \gamma_n)$. For notational simplicity, we denote the distribution of model $(-\mathbf{v}/2, \mathbf{v}/2, \mathbf{I}, \alpha)$ by $\mathbb{P}_{\mathbf{v}}$ and the product distribution of n i.i.d. samples by $\mathbb{P}_{\mathbf{v}}^n$. By the definition of the minimax risk in (2.4), we have

$$\sup_{\boldsymbol{\Sigma}} R_n^*[\mathcal{G}_0(\boldsymbol{\Sigma}), \mathcal{G}_1(\boldsymbol{\Sigma}; \gamma_n)] \geq \inf_{\phi} \left[\mathbb{P}_{\mathbf{0}}^n(\phi = 1) + \frac{1}{|\mathcal{H}(s)|} \sum_{\mathbf{v} \in \mathcal{H}(s)} \mathbb{P}_{\mathbf{v}}^n(\phi = 0) \right].$$

We thus reduce the minimax risk to the risk of a simple-against-simple hypothesis test where the alternative hypothesis corresponds to a uniform mixture of $\{\mathbb{P}_{\mathbf{v}} : \mathbf{v} \in \mathcal{H}(s)\}$. For notational simplicity, we define $\mathbb{P}_{\mathcal{H}}^n := 1/|\mathcal{H}(s)| \cdot \sum_{\mathbf{v} \in \mathcal{H}(s)} \mathbb{P}_{\mathbf{v}}^n$. By Neyman-Pearson Lemma, we have

$$R_n^*[\mathcal{G}_0, \mathcal{G}_1(\boldsymbol{\Sigma}; \gamma_n)] \geq 1 - \text{TV}(\mathbb{P}_{\mathbf{0}}^n, \mathbb{P}_{\mathcal{H}}^n).$$

Using Pinsker's inequality $\text{TV}(\mathbb{P}_{\mathbf{0}}^n, \mathbb{P}_{\mathcal{H}}^n) \leq \sqrt{D_{\chi^2}(\mathbb{P}_{\mathcal{H}}^n, \mathbb{P}_{\mathbf{0}}^n)}$, for showing $R_n^*[\mathcal{G}_0(\boldsymbol{\Sigma}), \mathcal{G}_1(\boldsymbol{\Sigma}; \gamma_n)] \rightarrow 1$ as n goes to infinity, it suffices to show that $D_{\chi^2}(\mathbb{P}_{\mathcal{H}}^n, \mathbb{P}_{\mathbf{0}}^n) = o(1)$. By calculation we have

$$\begin{aligned} D_{\chi^2}(\mathbb{P}_{\mathcal{H}}^n, \mathbb{P}_{\mathbf{0}}^n) &= \mathbb{E}_{\mathbb{P}_{\mathbf{0}}^n} \left\{ \left[\frac{d\mathbb{P}_{\mathcal{H}}^n}{d\mathbb{P}_{\mathbf{0}}^n}(Y, \mathbf{X}) - 1 \right]^2 \right\} = \mathbb{E}_{\mathbb{P}_{\mathbf{0}}^n} \left\{ \left[\frac{d\mathbb{P}_{\mathcal{H}}^n}{d\mathbb{P}_{\mathbf{0}}^n}(Y, \mathbf{X}) \right]^2 \right\} - 1 \\ &= \frac{1}{|\mathcal{H}(s)|^2} \sum_{\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{H}(s)} \mathbb{E}_{\mathbb{P}_{\mathbf{0}}^n} \left[\frac{d\mathbb{P}_{\mathbf{v}_1}^n d\mathbb{P}_{\mathbf{v}_2}^n}{d\mathbb{P}_{\mathbf{0}}^n d\mathbb{P}_{\mathbf{0}}^n}(Y, \mathbf{X}) \right] - 1 \\ &= \frac{1}{|\mathcal{H}(s)|^2} \sum_{\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{H}(s)} \left\{ \mathbb{E}_{\mathbb{P}_{\mathbf{0}}^n} \left[\frac{d\mathbb{P}_{\mathbf{v}_1}^n d\mathbb{P}_{\mathbf{v}_2}^n}{d\mathbb{P}_{\mathbf{0}}^n d\mathbb{P}_{\mathbf{0}}^n}(Y, \mathbf{X}) \right] \right\}^n - 1. \end{aligned} \quad (\text{A.2})$$

We utilize the following lemma to obtain an upper bound for the last term of (A.2). See §B.1 for the proof.

Lemma A.1. *For any $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{H}(s)$, we have*

$$\mathbb{E}_{\mathbb{P}_{\mathbf{0}}^n} \left[\frac{d\mathbb{P}_{\mathbf{v}_1}^n}{d\mathbb{P}_{\mathbf{0}}^n} \frac{d\mathbb{P}_{\mathbf{v}_2}^n}{d\mathbb{P}_{\mathbf{0}}^n}(Y, \mathbf{X}) \right] = \cosh(\langle \mathbf{v}_1, \mathbf{v}_2 \rangle / 2) + \alpha^2 \sinh(\langle \mathbf{v}_1, \mathbf{v}_2 \rangle / 2).$$

By Lemma A.1, we have

$$\begin{aligned} D_{\chi^2}(\mathbb{P}_{\mathcal{H}}^n, \mathbb{P}_{\mathbf{0}}^n) &= \frac{1}{|\mathcal{H}(s)|^2} \sum_{\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{H}(s)} [\cosh(1/2 \cdot \langle \mathbf{v}_1, \mathbf{v}_2 \rangle) + \alpha^2 \sinh(1/2 \cdot \langle \mathbf{v}_1, \mathbf{v}_2 \rangle)]^n - 1. \end{aligned} \quad (\text{A.3})$$

We define $\mathcal{C} := \{\mathcal{S} \subseteq [d] : |\mathcal{S}| = s\}$, and let $\mathbb{U}_{\mathcal{C}}$ be the uniform distribution over \mathcal{C} . Let $\mathcal{S}_1, \mathcal{S}_2 \sim \mathbb{U}_{\mathcal{C}}$ be two independent random sets. Then by (A.3), we have

$$D_{\chi^2}(\mathbb{P}_{\mathcal{H}}^n, \mathbb{P}_{\mathbf{0}}^n) = \mathbb{E}_{\mathcal{S}_1, \mathcal{S}_2} [\cosh(\beta^2/2 \cdot |\mathcal{S}_1 \cap \mathcal{S}_2|) + \alpha^2 \sinh(\beta^2/2 \cdot |\mathcal{S}_1 \cap \mathcal{S}_2|)]^n - 1.$$

We use the next lemma, proved in §B.2, to bound the above right-hand side.

Lemma A.2. *For any $x \geq 0$ and $v \in [0, 1]$, we have*

$$\cosh(x) + v \sinh(x) \leq \exp(2vx) \vee \cosh(2x). \quad (\text{A.4})$$

Proceeding with this result and letting random variable $Z \sim |\mathcal{S}_1 \cap \mathcal{S}_2|$, we have

$$\begin{aligned} D_{\chi^2}(\mathbb{P}_{\mathcal{H}}^n, \mathbb{P}_{\mathbf{0}}^n) &\leq \mathbb{E}_Z [\exp(\alpha^2 \beta^2 Z) \vee \cosh(\beta^2 Z)]^n - 1 \\ &= \mathbb{E}_Z [\exp(n\alpha^2 \beta^2 Z) \vee \cosh(\beta^2 Z)^n] - 1 \\ &= \mathbb{E}_Z \{ \exp(n\alpha^2 \beta^2 Z) \vee \mathbb{E}_U [\exp(\beta^2 Z U)] \} - 1, \end{aligned} \quad (\text{A.5})$$

where in the last step, we introduce a random variable U that is the summation of n independent Rademacher random variables over $\{-1, 1\}$. Then we have $\cosh(\beta^2 Z)^n = \mathbb{E}_U[\exp(\beta^2 Z U)]$. By (A.5), we have

$$\begin{aligned} D_{\chi^2}(\mathbb{P}_{\mathcal{H}}^n, \mathbb{P}_{\mathbf{0}}^n) &\leq \mathbb{E}_Z \mathbb{E}_U [\exp(n\alpha^2 \beta^2 Z) \vee \exp(\beta^2 Z U)] - 1 \\ &= \mathbb{E}_U \mathbb{E}_Z \{ \exp(n\alpha^2 \beta^2) \vee \exp(\beta^2 U) \}^Z - 1 \\ &\leq \mathbb{E}_U \left\{ \sup_{\mathcal{S}_1 \in \mathcal{C}} \mathbb{E}_{\mathcal{S}_2} [\exp(n\alpha^2 \beta^2) \vee \exp(\beta^2 U)]^{|\mathcal{S}_1 \cap \mathcal{S}_2|} \right\} - 1. \end{aligned} \quad (\text{A.6})$$

Now we turn to bound the expectation over \mathcal{S}_2 in (A.6). For any fixed \mathcal{S}_1 , we have

$$|\mathcal{S}_1 \cap \mathcal{S}_2| = \sum_{i \in \mathcal{S}_1} V_i,$$

where V_i is binary random variable that indicates whether $i \in \mathcal{S}_2$. It is known that V_1, \dots, V_d are negative associated. Hence we have

$$\begin{aligned} \mathbb{E}_{\mathcal{S}_2} [\exp(n\alpha^2 \beta^2) \vee \exp(\beta^2 U)]^{|\mathcal{S}_1 \cap \mathcal{S}_2|} &\leq \prod_{i \in \mathcal{S}_1} \mathbb{E}_{V_i} [\exp(n\alpha^2 \beta^2) \vee \exp(\beta^2 U)]^{V_i} \\ &= \{1 + s/d \cdot [\exp(n\alpha^2 \beta^2) \vee \exp(\beta^2 U) - 1]\}^s. \end{aligned} \quad (\text{A.7})$$

Plugging (A.7) into (A.6) and expanding the polynomial term, we have

$$\begin{aligned} D_{\chi^2}(\mathbb{P}_{\mathcal{H}}^n, \mathbb{P}_{\mathbf{0}}^n) &\leq \sum_{k=1}^s \binom{s}{k} \cdot (s/d)^k \cdot \mathbb{E}_U [\exp(n\alpha^2 \beta^2) \vee \exp(\beta^2 U) - 1]^k \\ &= \sum_{k=1}^s \binom{s}{k} \cdot (s/d)^k \cdot \left([\exp(n\alpha^2 \beta^2) - 1]^k \cdot \mathbb{P}(U < n\alpha^2) \right. \\ &\quad \left. + \mathbb{E}_U \left\{ [\exp(\beta^2 U) - 1]^k \mid U \geq \alpha^2 n \right\} \cdot \mathbb{P}(U \geq n\alpha^2) \right), \\ &\leq T_1 + T_2, \end{aligned}$$

where T_1 and T_2 are defined as

$$\begin{aligned} T_1 &:= \sum_{k=1}^s \binom{s}{k} \cdot (s/d)^k \cdot [\exp(n\alpha^2 \beta^2) - 1]^k \\ T_2 &:= \sum_{k=1}^s \binom{s}{k} \cdot (s/d)^k \cdot \mathbb{E}_U \left\{ [\exp(\beta^2 U) - 1]^k \mid U \geq 0 \right\} \cdot \mathbb{P}(U \geq 0). \end{aligned}$$

It remains to bound T_1 and T_2 respectively.

Bounding T_1 . Under condition $s\beta^2 = \gamma_n = o(1/\alpha^2 \cdot s \log d/n)$, we have $\beta^2 = o(1/\alpha^2 \cdot \log d/n)$. Hence, for any small constant $C > 0$, we have $\beta^2 \leq C \cdot 1/\alpha^2 \cdot \log d/n$ when n is sufficiently large. Note that we assume $s = o(d^{1/2-\delta})$ for some fixed constant $\delta > 0$. Then we have

$$\begin{aligned} T_1 &\leq \sum_{k=1}^s \binom{s}{k} \cdot (s/d)^k \cdot \exp(\alpha^2 \beta^2 nk) \leq \sum_{k=1}^s [s^2 e/(kd)]^k \cdot \exp(\alpha^2 \beta^2 nk) \\ &\leq \sum_{k=1}^s [s^2 e/(kd)]^k \cdot \exp(Ck \log d) = \sum_{k=1}^s (s^2 e/k \cdot d^{C-1})^k \leq \sum_{k=1}^s (e/k \cdot d^{C-2\delta})^k, \end{aligned}$$

where the second step follows from the fact that $\binom{s}{k} \leq (es/k)^k$. Note that C is chosen arbitrarily, hence we can always choose $C \leq \delta$. It implies that $e/k \cdot d^{C-2\delta} = o(1)$. We thus conclude $T_1 = o(1)$.

Bounding T_2 . For term T_2 , we observe that

$$\begin{aligned} T_2 &\leq \sum_{k=1}^s (e/k \cdot s^2/d)^k \cdot \mathbb{E}_U \left\{ [\exp(\beta^2|U|) - 1]^k \right\} \\ &\leq \sum_{k=1}^s (e/k \cdot s^2/d)^k \cdot \mathbb{E}_U \left[(\beta^2|U|)^k + \exp(\beta^2 k|U|) \cdot \mathbf{1}(\beta^2|U| \geq 1) \right] \\ &\leq T_3 + T_4, \end{aligned}$$

where T_3 and T_4 are defined as

$$\begin{aligned} T_3 &:= \sum_{k=1}^s \mathbb{E}_U (e/k \cdot s^2 \beta^2/d \cdot |U|)^k, \\ T_4 &:= \sum_{k=1}^s (e/k \cdot s^2/d)^k \cdot \mathbb{E}_U \left[\exp(\beta^2 k|U|) \cdot \mathbf{1}(\beta^2|U| \geq 1) \right]. \end{aligned}$$

Note that U is summation of n i.i.d. centered sub-Gaussian random variables U_i each with Orlicz ψ_2 -norm equal to one. Therefore, U is also centered sub-Gaussian random variable with $\|U\|_{\psi_2} \leq C\sqrt{n}$ for some constant C . Thus it holds that

$$\mathbb{E}(|U|^k) \leq (\sqrt{k} \cdot \|U\|_{\psi_2})^k \leq (C\sqrt{nk})^k.$$

Hence for term T_3 , we have

$$T_3 \leq \sum_{k=1}^s \left[Ces^2 \beta^2 \sqrt{n}/(\sqrt{kd}) \right]^k,$$

Under the condition $s\beta^2 = o(\sqrt{s \log d/n})$, we have

$$Ces^2 \beta^2 \sqrt{n}/(\sqrt{kd}) = o\left(s\sqrt{s \log d/d}\right).$$

Since $s = o(\sqrt{d})$, we have $s\sqrt{s \log d/d} = o(1)$, which implies $T_3 = o(1)$.

To obtain an upper bound for term T_4 , we let $W = \beta^2 U$. So W is centered sub-Gaussian with Orlicz norm $c\beta^2 \sqrt{n}$. Computing integral by parts, we have

$$\mathbb{E}_U \left[\exp(\beta^2 k|U|) \cdot \mathbf{1}(\beta^2|U| \geq 1) \right] = e^k \cdot \mathbb{P}(|W| \geq 1) + \int_{w=1}^{\infty} k e^{wk} \cdot \mathbb{P}(|W| \geq w) dw. \quad (\text{A.8})$$

Using the property of sub-Gaussianity, we have $\mathbb{P}[W \geq t] \leq C_1 \exp[-C_2 t^2/(\beta^2 \sqrt{n})^2]$ for some absolute constants $C_1, C_2 > 0$. Proceeding with (A.8) and using shorthand $\sigma = \beta^2 \sqrt{n}$, we obtain

$$\begin{aligned} \mathbb{E}_U \left[\exp(\beta^2 k|U|) \cdot \mathbf{1}(\beta^2|U| \geq 1) \right] &\leq C_1 e^k e^{-C_2/\sigma^2} + C_1 k \int_{w=1}^{\infty} e^{wk} e^{-C_2 w^2/\sigma^2} dw \\ &= C_1 e^k e^{-C_2/\sigma^2} + C_1 k e^{k^2 \sigma^2/(4C_2)} \int_{w=1}^{\infty} e^{-\frac{C_2}{\sigma^2} (w - \frac{k\sigma^2}{2C_2})^2} dw \leq C_1 e^k + C_3 k e^{k^2 \sigma^2/(4C_2)} \sigma, \end{aligned}$$

where C_3 is a constant that depends on C_1 and C_2 . Thus we have

$$T_4 \leq \underbrace{\sum_{k=1}^s C_1 [s^2 e^2/(kd)]^k}_{T_5} + \underbrace{\sum_{k=1}^s C_3 \sigma k [s^2 e^2/(kd) \cdot \exp(k/4 \cdot \sigma^2/C_2)]^k}_{T_6}. \quad (\text{A.9})$$

Note that $s^2/d = o(1)$, we thus have $T_5 = o(1)$. Under condition $s\beta^2 = o(\sqrt{s \log d/n})$, for any small constant $C > 0$, when n is large enough, we have

$$\exp(k/4 \cdot \sigma^2/C_2) \leq \exp(Ck \log d/s) \leq \exp(C \log d) \leq d^C.$$

Plugging (A.9) into T_6 and using $s^2 = o(d^{1-2\delta})$, we have that each term in the summation is less than

$$T_6 \leq \sum_{k=1}^s \sigma k [e^2/(kd^{2\delta-C})]^k \lesssim \sum_{k=1}^s k \sqrt{\log d/s} \cdot [e^2/(d^{2\delta-C})]^k.$$

Since the constant C is chosen arbitrarily, we have $T_6 = o(1)$. Accordingly, $T_4 = o(1)$ and $T_2 = o(1)$.

Finally, combining everything together, we have $D_{\chi^2}(\mathbb{P}_{\mathcal{H}}^n, \mathbb{P}_{\mathbf{0}}^n) = o(1)$, which completes the proof.

A.2 Proof of Theorem 3.2

We begin with some basic properties of sample sets $\{\mathbf{w}_i\}_{i=1}^n$ and $\{\mathbf{u}_i\}_{i=1}^n$. We introduce the random vector $\mathbf{W} := \mathbf{X} - \mathbf{X}'$ to capture the distribution of samples $\{\mathbf{w}_i\}_{i=1}^n$. Here \mathbf{X} follows the model given in (2.1)-(2.2), and \mathbf{X}' is an independent copy of \mathbf{X} . We note that the marginal distribution of \mathbf{X} is given by $1/2 \cdot \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) + 1/2 \cdot \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$. Thus \mathbf{W} follows a mixture distribution

$$\mathbf{W} \sim 1/2 \cdot \mathcal{N}(\mathbf{0}, 2\boldsymbol{\Sigma}) + 1/4 \cdot \mathcal{N}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0, 2\boldsymbol{\Sigma}) + 1/4 \cdot \mathcal{N}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1, 2\boldsymbol{\Sigma}). \quad (\text{A.10})$$

Moreover, conditioning on the observed label Y , the distribution of \mathbf{X} is given by

$$\mathbf{X}|Y=0 \sim (1+\alpha)/2 \cdot \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) + (1-\alpha)/2 \cdot \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}), \quad (\text{A.11})$$

$$\mathbf{X}|Y=1 \sim (1+\alpha)/2 \cdot \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + (1-\alpha)/2 \cdot \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}). \quad (\text{A.12})$$

We introduce a random vector $\mathbf{U} := \mathbf{X}^{(1)} - \mathbf{X}^{(0)}$ that corresponds to samples $\{\mathbf{u}_i\}_{i=1}^n$. Here random vectors $\mathbf{X}^{(0)}$ and $\mathbf{X}^{(1)}$ are independent and have distributions given in (A.11), (A.12), respectively. The distribution of \mathbf{U} is given by

$$\mathbf{U} \sim (1+\alpha)^2/4 \cdot \mathcal{N}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0, 2\boldsymbol{\Sigma}) + (1-\alpha)^2/2 \cdot \mathcal{N}(\mathbf{0}, 2\boldsymbol{\Sigma}) + (1-\alpha)^2/4 \cdot \mathcal{N}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1, 2\boldsymbol{\Sigma}). \quad (\text{A.13})$$

Now we turn to prove Theorem 3.2. It suffices to prove this result by bounding type-I and type-II errors separately. In the end, we will show that

$$\sup_{\boldsymbol{\theta} \in \mathcal{G}_0(\boldsymbol{\Sigma})} \mathbb{P}_{\boldsymbol{\theta}}^n(\phi = 1) \leq 4d^{-1} \quad \text{and} \quad \sup_{\boldsymbol{\theta} \in \mathcal{G}_1(\boldsymbol{\Sigma}; \gamma_n)} \mathbb{P}_{\boldsymbol{\theta}}^n(\phi = 0) \leq 16d^{-1}.$$

Type-I error. Under the null hypothesis $\boldsymbol{\theta} \in \mathcal{G}_0(\boldsymbol{\Sigma})$, (A.10) and (A.13) reduce to

$$\mathbf{W} \sim \mathcal{N}(\mathbf{0}, 2\boldsymbol{\Sigma}), \quad \mathbf{U} \sim \mathcal{N}(\mathbf{0}, 2\boldsymbol{\Sigma}).$$

To bound the type-I error of function ϕ_1 , we first note that

$$\frac{1}{n} \sum_{i=1}^n (\mathbf{v}^\top \boldsymbol{\Sigma}^{-1} \mathbf{w}_i)^2 = \mathbf{v}^\top \widehat{\boldsymbol{\Sigma}}_W \mathbf{v},$$

where we let $\widehat{\boldsymbol{\Sigma}}_W := 1/n \cdot \sum_{i=1}^n \boldsymbol{\Sigma}^{-1} \mathbf{w}_i \mathbf{w}_i^\top \boldsymbol{\Sigma}^{-1}$, i.e., an empirical covariance matrix of random vector $\boldsymbol{\Sigma}^{-1} \mathbf{W} \sim \mathcal{N}(\mathbf{0}, 2\boldsymbol{\Sigma}^{-1})$. For any matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ and $\mathcal{S} \subseteq [d]$, we let $[\mathbf{A}]_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ be the submatrix of \mathbf{A} , which contains the entries with row and column indices in \mathcal{S} . By standard tail bound of Gaussian covariance estimation (see Lemma C.2), for any fixed $\mathcal{S} \in [d]$ with $|\mathcal{S}| = s$, and any $\epsilon \in (0, 1)$, when $n \geq Cs/\epsilon^2$ for some constant C , we have

$$\mathbb{P}_{\boldsymbol{\theta}}^n \left[\|\widehat{\boldsymbol{\Sigma}}_W - 2\boldsymbol{\Sigma}^{-1}\|_{\mathcal{S}} \geq 2\epsilon \|\boldsymbol{\Sigma}^{-1}\|_{\mathcal{S}} \right] \leq 2e^{-n}. \quad (\text{A.14})$$

Note that $\|\boldsymbol{\Sigma}^{-1}\|_{\mathcal{S}} \leq \|\boldsymbol{\Sigma}^{-1}\|_2$ for all $\mathcal{S} \subseteq [d]$. By taking union bound over all subsets with size s in $[d]$, we have

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\theta}}^n \left[\sup_{\mathcal{S} \in [d], |\mathcal{S}|=s} \|\widehat{\boldsymbol{\Sigma}}_W - 2\boldsymbol{\Sigma}^{-1}\|_{\mathcal{S}} \geq 2\epsilon \|\boldsymbol{\Sigma}^{-1}\|_2 \right] &\leq 2 \binom{d}{s} e^{-n} \\ &\stackrel{(a)}{\leq} 2 \exp[-n + s \log(ed/s)] \stackrel{(b)}{\leq} 2[s/(ed)]^s \leq 2d^{-1}. \end{aligned}$$

Here step (a) follows from the fact that $\binom{d}{s} \leq (ed/s)^s$ and step (b) follows from the assumption that $n \geq 2s \log(ed/s)$. In the last step we use the fact that function $f(s) = (s/d)^s$ is monotonically decreasing for $s \in [1, d/e]$. We set $\epsilon = \sqrt{s \log(ed/s)/n}$. Under condition $n \geq 2s \log(ed/s)$, we have $\epsilon < 1$. Moreover, when $s \leq C'd$ for sufficiently small constant C' that depends on C , we have $n \geq Cs/\epsilon^2$. Therefore, such value of ϵ leads to (A.14). Thus we conclude that

$$\mathbb{P}_{\boldsymbol{\theta}}^n \left[\frac{\mathbf{v}^\top \widehat{\boldsymbol{\Sigma}}_W \mathbf{v} - 2\mathbf{v}^\top \boldsymbol{\Sigma}^{-1} \mathbf{v}}{2\mathbf{v}^\top \boldsymbol{\Sigma}^{-1} \mathbf{v}} \geq \sqrt{\frac{s \log(ed/s)}{n}} \cdot \frac{\|\boldsymbol{\Sigma}^{-1}\|_2}{\mathbf{v}^\top \boldsymbol{\Sigma}^{-1} \mathbf{v}}, \text{ for all } \mathbf{v} \in \mathcal{B}_2(s) \right] \leq 2d^{-1}$$

Note that $\|\Sigma^{-1}\|_2/(\mathbf{v}^\top \Sigma^{-1} \mathbf{v}) \leq \|\Sigma^{-1}\|_2 \|\Sigma\|_2 = \kappa$. Our choice of τ_1 ensures the type-I error of ϕ_1 does not exceed $2d^{-1}$.

Now we turn to analyze the performance of ϕ_2 . Recall that ϕ_1 simply selects the coordinate of $\bar{\mathbf{u}} := 1/n \cdot \sum_{i=1}^n \mathbf{u}_i$ that has the largest magnitude (scaled with $\text{diag}(\Sigma)^{-1/2}$) and compare it with τ_2 . It suffices to show all coordinates are well bounded around 0 under null hypothesis. Denote the j -th coordinate of $\bar{\mathbf{u}}$ by \bar{u}_j . Denote the j -th diagonal term of Σ by σ_j . We have $\bar{u}_j \sim \mathcal{N}(0, 2\sigma_j/n)$. Recall that for standard normal random variable X , we have

$$\mathbb{P}(|X| \geq t) \leq 2 \exp(-t^2/2) \text{ for any } t \geq 1. \quad (\text{A.15})$$

Using this property and taking union bound over $j \in [d]$, we have

$$\mathbb{P}_\theta^n \left(\sup_{j \in [d]} |\bar{u}_j|/\sqrt{\sigma_j} \geq 8 \log d/n \right) \leq 2d \cdot \exp(-2 \log d) = 2d^{-1}.$$

Accordingly, our choice of τ_2 can ensure type-I error of ϕ_2 is controlled within $2d^{-1}$.

Type-II error. Under the alternative hypothesis $\theta \in \mathcal{G}_1(\Sigma; \gamma_n)$. Note that $\phi = 0$ if and only if $\phi_1 = 0$ and $\phi_2 = 0$. Thus, for any $\theta \in \mathcal{G}_1(\Sigma; \gamma_n)$, we have

$$\mathbb{P}_\theta^n(\phi = 0) = \mathbb{P}_\theta^n(\phi_1 = 0 \cap \phi_2 = 0) \leq \mathbb{P}_\theta^n(\phi_1 = 0) \wedge \mathbb{P}_\theta^n(\phi_2 = 0). \quad (\text{A.16})$$

We assume $\gamma_n \geq C\kappa[\sqrt{s \log d/n} \vee (1/\alpha^2 \cdot s \log d/n)]$. It suffices to bound the type-II error by considering these two cases: (i) when $\gamma_n \gtrsim \kappa \sqrt{s \log d/n}$, we show that $\mathbb{P}_\theta^n(\phi_1 = 0) \leq 16d^{-1}$; (ii) when $\gamma_n \gtrsim \kappa/\alpha^2 \cdot s \log d/n$ and $16/\alpha^2 \cdot s \log d/n \leq \sqrt{s \log d/n}$, we show $\mathbb{P}_\theta^n[\phi_2 = 0] \leq 7d^{-1}$.

Case (i). Now we consider the first case. We denote $\Delta\boldsymbol{\mu} := \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$. Let $\mathbf{v}^* := \Delta\boldsymbol{\mu}/\|\Delta\boldsymbol{\mu}\|_2$. Since $\mathbf{v}^* \in \mathcal{B}_2(s)$, we have

$$\sup_{\mathbf{v} \in \mathcal{B}_2(s)} \frac{\mathbf{v}^\top \widehat{\Sigma}_W \mathbf{v}}{2\mathbf{v}^\top \Sigma^{-1} \mathbf{v}} \geq \frac{\mathbf{v}^{*\top} \widehat{\Sigma}_W \mathbf{v}^*}{2\mathbf{v}^{*\top} \Sigma^{-1} \mathbf{v}^*}.$$

It remains to show the right hand side is larger than $1 + \tau_1$ with high probability. Note that

$$\mathbf{v}^{*\top} \widehat{\Sigma}_W \mathbf{v}^* = \frac{1}{n} \sum_{i=1}^n (\mathbf{v}^{*\top} \Sigma^{-1} \mathbf{w}_i)^2.$$

We define a random variable $\widetilde{W} := \mathbf{v}^{*\top} \Sigma^{-1} \mathbf{W}$, whose probability distribution is given by

$$1/2 \cdot \mathcal{N}(0, \nu) + 1/4 \cdot \mathcal{N}(m, \nu) + 1/4 \cdot \mathcal{N}(-m, \nu), \quad (\text{A.17})$$

where we define $m := \rho(\theta)/\|\Delta\boldsymbol{\mu}\|_2$ and $\nu := 2\rho(\theta)/\|\Delta\boldsymbol{\mu}\|_2^2$. Recall that $\rho(\theta) := \Delta\boldsymbol{\mu}^\top \Sigma^{-1} \Delta\boldsymbol{\mu}$. Let $\tilde{w}_i := \mathbf{v}^{*\top} \Sigma^{-1} \mathbf{w}_i$. Due to the mixture structure (A.17), we can thus cluster $\{\tilde{w}_i\}_{i=1}^n$ into three groups $\{\tilde{w}_i^{(k)}\}_{i=1}^{n_k}$, $k \in \{1, 2, 3\}$, based on the latent labels. The k -th group corresponds to the k -th term in (A.17). Note that $\mathbb{E}(n_1) = n/2$, $\mathbb{E}(n_2) = \mathbb{E}(n_3) = n/4$. Define event \mathcal{E}_1 as

$$\mathcal{E}_1 := \{|n_1 - n/2| \leq 1/8 \cdot n, |n_2 - n/4| \leq 1/8 \cdot n, |n_3 - n/4| \leq 1/8 \cdot n\}. \quad (\text{A.18})$$

By Hoeffding's inequality, we have $\mathbb{P}(\mathcal{E}_1) \geq 1 - 6 \exp(-n^2/32)$.

From now on, we condition on event \mathcal{E}_1 . By the standard χ^2 -tail bound (Lemma C.1), for any $t \in (0, 1)$ and $k \in \{1, 2, 3\}$, we have

$$\mathbb{P}_\theta^n \left(\left| \sum_{i=1}^{n_k} (\tilde{w}_i^{(k)} - m_k)^2 - n_k \nu \right| \geq n_k \nu t \right) \leq 2e^{-n_k t^2/8} \leq 2e^{-nt^2/64}, \quad (\text{A.19})$$

where $m_1 = 0, m_2 = -m_3 = m$. Moreover, using tail bound of Gaussian (A.15), for $t' \geq 1/\sqrt{n_k}$ and $k = 2, 3$,

$$\mathbb{P}_\theta^n \left(\left| \sum_{i=1}^{n_k} \tilde{w}_i^{(k)} - n_k m_k \right| \geq n_k \sqrt{\nu} t' \right) \leq 2e^{-n_k t'^2/2} \leq 2e^{-nt'^2/16}. \quad (\text{A.20})$$

Excluding the small chance events in (A.19) and (A.20), we find that

$$\begin{aligned}
\sum_{i=1}^n \tilde{w}_i^2 &= \sum_{k=1}^3 \sum_{i=1}^{n_k} (\tilde{w}_i^{(k)} - m_k)^2 + 2 \sum_{k=2}^3 \sum_{i=1}^{n_k} m_k \tilde{w}_i^{(k)} - (n_2 + n_3) m^2 \\
&\geq n\nu(1-t) + 2 \sum_{k=2}^3 \sum_{i=1}^{n_k} m_k \tilde{w}_i^{(k)} - (n_2 + n_3) m^2 \\
&\geq n\nu(1-t) + (n_2 + n_3) m^2 - 2(n_2 + n_3) \sqrt{\nu} t' m \\
&\geq n\nu(1-t) + 1/4 \cdot n m^2 - 3/2 \cdot n \sqrt{\nu} t' m,
\end{aligned}$$

where the last step follows from (A.18). Note that $2\mathbf{v}^{*\top} \Sigma^{-1} \mathbf{v}^* = \nu$. We thus have

$$\begin{aligned}
\frac{\mathbf{v}^{*\top} \widehat{\Sigma}_W \mathbf{v}^*}{2\mathbf{v}^{*\top} \Sigma^{-1} \mathbf{v}^*} - 1 &= \frac{\sum_{i=1}^n \tilde{w}_i^2}{2n\mathbf{v}^{*\top} \Sigma^{-1} \mathbf{v}^*} - 1 \geq \frac{m^2}{4\nu} - t - \frac{3mt'}{2\sqrt{\nu}} \\
&= 1/8 \cdot \rho(\boldsymbol{\theta}) - t - 3t'/4 \cdot \sqrt{2\rho(\boldsymbol{\theta})}. \tag{A.21}
\end{aligned}$$

Now we choose $t = t' = 8\sqrt{s \log(ed/s)/n}$, which is less than one under condition $n \geq 64s \log(ed/s)$. When $\rho(\boldsymbol{\theta}) \geq C\kappa\sqrt{s \log(ed/s)/n}$ for sufficiently large constant C , we can have $t \leq \rho(\boldsymbol{\theta})/32$ and $t' \leq \sqrt{t'} \leq \sqrt{\rho(\boldsymbol{\theta})}/48$. Accordingly, proceeding with (A.21) gives

$$1/2 \cdot \mathbf{v}^{*\top} \widehat{\Sigma}_W \mathbf{v}^* / \mathbf{v}^{*\top} \Sigma^{-1} \mathbf{v}^* - 1 \geq 1/16 \cdot \rho(\boldsymbol{\theta}) \geq \tau_1.$$

Plugging the value of t, t' into the tail bounds in (A.19) (A.20) and using the probability of event \mathcal{E}_1 , we have the type-II error of ϕ_1 is most $10d^{-1} + 6e^{-n^2/32} \leq 16d^{-1}$.

Case (ii). Now we turn to analyze the performance of ϕ_2 . We introduce shorthands $\tilde{\boldsymbol{\mu}} := \text{diag}(\Sigma)^{-1/2} \Delta \boldsymbol{\mu}$ and $\Lambda := \text{diag}(\Sigma)^{1/2}$. Then it holds that

$$\begin{aligned}
\rho(\boldsymbol{\theta}) &= \Delta \boldsymbol{\mu}^\top \Sigma^{-1} \Delta \boldsymbol{\mu} = \Delta \boldsymbol{\mu}^\top \Lambda^{-1} \Lambda \Sigma^{-1} \Lambda \Lambda^{-1} \Delta \boldsymbol{\mu} \leq \|\tilde{\boldsymbol{\mu}}\|_2^2 \|\Lambda \Sigma^{-1} \Lambda\|_{op} \\
&\leq \|\tilde{\boldsymbol{\mu}}\|_2^2 \|\Lambda\|_2^2 \|\Sigma^{-1}\|_2 \leq \kappa \|\tilde{\boldsymbol{\mu}}\|_2^2,
\end{aligned}$$

where the last step follows from the fact that $\|\text{diag}(\Sigma)\|_2 \leq \|\Sigma\|_2$. Suppose the j -th coordinate of $\tilde{\boldsymbol{\mu}}$, denoted by β , has largest magnitude. Since $\|\tilde{\mathbf{u}}\|_2^2 \leq s\beta^2$, we have $\beta^2 \geq \rho(\boldsymbol{\theta})/(s\kappa)$. Under condition

$$\rho(\boldsymbol{\theta}) \geq \gamma_n \geq \frac{400\kappa s \log d}{\alpha^2 n},$$

we have

$$\beta \geq 20\sqrt{\log d/(\alpha^2 n)}. \tag{A.22}$$

Let $\mathbf{v}^* = \text{sign}(\beta) \cdot \mathbf{e}_j$. We have

$$\sup_{\mathbf{v} \in \mathcal{B}_2(1)} \langle \mathbf{v}, \Lambda^{-1} \tilde{\mathbf{u}} \rangle \geq \langle \mathbf{v}^*, \Lambda^{-1} \tilde{\mathbf{u}} \rangle = \left| \frac{1}{n} \sum_{i=1}^n \tilde{u}_{ij} \right|,$$

where we denote the j -th coordinate of $\Lambda^{-1} \mathbf{u}_i$ by \tilde{u}_{ij} .

Let U_j be the j -th coordinate of \mathbf{U} . Note that $\{\tilde{u}_{ij}\}_{i=1}^n$ are i.i.d. samples of $U_j/\sqrt{\sigma_j}$. Recall that σ_j is the j -th diagonal term of Σ . According to (A.13), $U_j/\sqrt{\sigma_j}$ has the mixture distribution

$$(1 + \alpha)^2/4 \cdot \mathcal{N}(\beta, 2) + (1 - \alpha^2)/2 \cdot \mathcal{N}(0, 2) + (1 - \alpha)^2/4 \cdot \mathcal{N}(-\beta, 2). \tag{A.23}$$

We can cluster these samples into three groups $\{\tilde{u}_{ij}^{(k)}\}_{i=1}^{n_k}, k \in \{1, 2, 3\}$ based on latent labels, where k -th group corresponds to the k -th term in (A.23). Using tail bound of Gaussian (A.15), we have for $t \geq 1$ and $k \in \{1, 2, 3\}$,

$$\mathbb{P}_{\boldsymbol{\theta}}^n \left(\left| \sum_{i=1}^{n_k} \tilde{u}_{ij}^{(k)} - n_k m_k \right| \geq \sqrt{2n_k t} \right) \leq 2e^{-t^2/2},$$

where $m_1 = -m_3 = \beta, m_2 = 0$. Therefore, with probability at least $1 - 6e^{-t^2/2}$, it holds that

$$\left| \frac{1}{n} \sum_{i=1}^n \tilde{u}_{ij} - \frac{(n_1 - n_3)\beta}{n} \right| \leq t \cdot \sum_{k=1}^3 \sqrt{\frac{2n_k}{n^2}} \leq \frac{5t}{\sqrt{n}}. \tag{A.24}$$

It remains to bound $n_1 - n_3$. Note that $n_1 - n_3$ is a summation of n i.i.d. random variables V_i satisfying $\mathbb{P}(V_i = 1) = (1 + \alpha)^2/4$, $\mathbb{P}(V_i = 0) = (1 - \alpha^2)/2$, and $\mathbb{P}(V_i = -1) = (1 - \alpha)^2/4$. Then V_i has mean α , variance $(1 - \alpha^2)/2 \leq 1 - \alpha$, and $|V_i - \mathbb{E}(V_i)| \leq 2$. By Bernstein's inequality, we have that for $t' > 0$,

$$\mathbb{P}(|n_1 - n_3 - \alpha n| \geq t') \leq \exp\left[-\frac{t'^2}{2(1 - \alpha)n + 4t'/3}\right].$$

Choosing $t' = \alpha n/2$, we thus have

$$\mathbb{P}(|n_1 - n_3 - \alpha \cdot n| \geq \alpha n/2) \leq \exp\left[-\frac{\alpha^2 n}{8(1 - \alpha) + 8\alpha/3}\right] \leq \exp(-\alpha^2 n/8) \leq d^{-1}, \quad (\text{A.25})$$

where the last step follows from condition $8s \log d/(\alpha^2 n) \leq \sqrt{s \log(ed/s)/n} \leq 1$. Combining (A.24) and (A.25), we have that with high probability $1 - 6e^{-t^2/2} - d^{-1}$,

$$\left|1/n \cdot \sum_{i=1}^n \tilde{u}_{ij}\right| \geq \alpha\beta/2 - 5t/\sqrt{n} \geq 10\sqrt{\log d/n} - 5t/\sqrt{n} \geq \tau_2,$$

where the second step follows from (A.22) and the last inequality holds by setting $t = \sqrt{2 \log d}$, which gives the type-II error of ϕ_2 is at most $7d^{-1}$.

Using (A.16) and the conclusions in the above two cases, we thus show Type-II error of ϕ is at most $16d^{-1}$ and thus complete the proof.

A.3 Proof of Theorem 3.3

In this section, we prove the computational lower bound. We first show that the information-theoretic lower bound in (3.4) is a lower bound of the computationally tractable minimax rate. To see this, we consider the oracle r^* that returns sample average $n^{-1} \sum_{i=1}^n q(y_i, \mathbf{x}_i)$ for any query function q . As discussed in §2.2, Bernstein's inequality in (2.6) and uniform concentration of empirical process imply that $r^* \in \mathcal{R}[\xi, n, T_n, \eta(\mathcal{Q}_{\mathcal{A}})]$. In addition, every test function ϕ that is based on the responses of r^* is also a function of $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$. Thus combining (2.4) and (2.7), it holds that

$$\bar{R}_n^*(\mathcal{G}_0, \mathcal{G}_1; \mathcal{A}, r^*) \geq R_n^*(\mathcal{G}_0, \mathcal{G}_1).$$

Therefore, by Theorem 3.1, for any γ_n satisfying

$$\gamma_n = o\left[\sqrt{s \log d/n} \wedge (1/\alpha^2 \cdot s \log d/n)\right],$$

we have $\lim_{n \rightarrow \infty} \bar{R}_n^*[\mathcal{G}_0, \mathcal{G}_1(\gamma_n); \mathcal{A}, r^*] = 1$. Here the equality holds because a test based on purely random guess incurs risk one.

Based on this observation, to show Theorem 3.3, in the following, we assume that

$$\gamma_n = o\left[\sqrt{s^2/n} \wedge (1/\alpha^2 \cdot s/n)\right]. \quad (\text{A.26})$$

We show that under this assumption, there exists an oracle r such that the minimax testing risk is not negligible. Similar to the derivation of the information theoretical lower bound, we also focus on the restricted testing problem defined in (A.1). Following the same notations, we denote by $\mathbb{P}_{\mathbf{0}}$ the distribution of model $(\mathbf{0}, \mathbf{0}, \mathbf{I}, \alpha)$ and by $\mathbb{P}_{\mathbf{v}}$ the distribution of model $(-\mathbf{v}/2, \mathbf{v}/2, \mathbf{I}, \alpha)$ for all $\mathbf{v} \in \mathcal{H}(s) = \{\mathbf{u} \in \{0, \beta\}^d: \|\mathbf{u}\|_0 = s\}$. Here we assume that the SNR under H_1 satisfies $\beta^2 s = \gamma_n$. Moreover, we define $\bar{\mathbb{P}}_{\mathbf{0}}$ as the distribution of the random variables returned by the statistical query model under the null hypothesis H_0 and define $\bar{\mathbb{P}}_{\mathbf{v}}$ correspondingly. Then the minimax testing risk $\bar{R}_n^*(\mathcal{G}_0, \mathcal{G}_1; \mathcal{A}, r)$ defined in (2.7) is lower bounded by

$$\sup_{\Sigma} \bar{R}_n^*[\mathcal{G}_0(\Sigma), \mathcal{G}_1(\Sigma; \gamma_n); \mathcal{A}, r] \geq \inf_{\phi \in \mathcal{H}(\mathcal{A}, r)} \left[\bar{\mathbb{P}}_{\mathbf{0}}(\phi = 1) + \frac{1}{|\mathcal{H}(s)|} \sum_{\mathbf{v} \in \mathcal{H}(s)} \bar{\mathbb{P}}_{\mathbf{v}}(\phi = 0) \right].$$

The following lemma establishes a sufficient condition that any hypothesis test under the statistical query model is asymptotically powerless. See [27] and [8] for a proof.

Lemma A.3. For any algorithm $\mathcal{A} \in \mathcal{A}(T)$ and any query function $q \in \mathcal{Q}_{\mathcal{A}}$, we define

$$\begin{aligned} \mathcal{C}_1(q) &= \{\mathbf{v} \in \mathcal{H}(s) : \mathbb{E}_{\mathbb{P}_{\mathbf{v}}} [q(Y, \mathbf{X})] - \mathbb{E}_{\mathbb{P}_0} [q(Y, \mathbf{X})] > \tau_q(\mathbb{P}_{\mathbf{v}})\}, \\ \mathcal{C}_2(q) &= \{\mathbf{v} \in \mathcal{H}(s) : \mathbb{E}_{\mathbb{P}_0} [q(Y, \mathbf{X})] - \mathbb{E}_{\mathbb{P}_{\mathbf{v}}} [q(Y, \mathbf{X})] > \tau_q(\mathbb{P}_{\mathbf{v}})\}. \end{aligned}$$

Here $\tau_q(\mathbb{P}_{\mathbf{v}})$ is the tolerance parameter defined in (2.5) when $(Y, \mathbf{X}) \sim \mathbb{P}_{\mathbf{v}}$. Then if $T \cdot \sup_{q \in \mathcal{Q}_{\mathcal{A}}} (|\mathcal{C}_1(q)| + |\mathcal{C}_2(q)|) / |\mathcal{H}(s)| = o(1)$, there exists an oracle $r \in \mathcal{R}[\xi, n, T, \eta(\mathcal{Q}_{\mathcal{A}})]$ such that

$$\inf_{\phi \in \mathcal{H}(\mathcal{A}, r)} \left[\bar{\mathbb{P}}_0(\phi = 1) + \frac{1}{|\mathcal{H}(s)|} \sum_{\mathbf{v} \in \mathcal{H}(s)} \bar{\mathbb{P}}_{\mathbf{v}}(\phi = 0) \right] = 1.$$

By this lemma, we need to construct an upper bound for $\sup_{q \in \mathcal{Q}_{\mathcal{A}}} (|\mathcal{C}_1(q)| + |\mathcal{C}_2(q)|)$. In the sequel, we achieve this goal by studying the uniform mixture of $\{\mathbb{P}_{\mathbf{v}} : \mathbf{v} \in \mathcal{C}_{\ell}(q)\}$ for $\ell \in \{1, 2\}$. Specifically, we define

$$\mathbb{P}_{\mathcal{C}_1(q)} = \frac{1}{|\mathcal{C}_1(q)|} \sum_{\mathbf{v} \in \mathcal{C}_1(q)} \mathbb{P}_{\mathbf{v}} \quad \text{and} \quad \mathbb{P}_{\mathcal{C}_2(q)} = \frac{1}{|\mathcal{C}_2(q)|} \sum_{\mathbf{v} \in \mathcal{C}_2(q)} \mathbb{P}_{\mathbf{v}}. \quad (\text{A.27})$$

The following lemma, obtained from [8], establishes an upper bound for the χ^2 -divergence between $\mathbb{P}_{\mathcal{C}_{\ell}(q)}$ and \mathbb{P}_0 .

Lemma A.4. For $\ell \in \{1, 2\}$ we define

$$\bar{\mathcal{C}}_{\ell}(q, \mathbf{v}) = \operatorname{argmax}_c \left\{ \frac{1}{|c|} \sum_{\mathbf{v}' \in c \subseteq \mathcal{H}(s)} \mathbb{E}_{\mathbb{P}_0} \left[\frac{d\mathbb{P}_{\mathbf{v}}}{d\mathbb{P}_0} \frac{d\mathbb{P}_{\mathbf{v}'}}{d\mathbb{P}_0} (Y, \mathbf{X}) \right] - 1 \mid |c| = |\mathcal{C}_{\ell}(q)| \right\}. \quad (\text{A.28})$$

Then the χ^2 -divergence between $\mathbb{P}_{\mathcal{C}_{\ell}(q)}$ and \mathbb{P}_0 is bounded by

$$D_{\chi^2}(\mathbb{P}_{\mathcal{C}_{\ell}(q)}, \mathbb{P}_0) \leq \sup_{\mathbf{v} \in \mathcal{C}_{\ell}(q)} \frac{1}{|\bar{\mathcal{C}}_{\ell}(q, \mathbf{v})|} \sum_{\mathbf{v}' \in \bar{\mathcal{C}}_{\ell}(q, \mathbf{v})} \mathbb{E}_{\mathbb{P}_0} \left[\frac{d\mathbb{P}_{\mathbf{v}}}{d\mathbb{P}_0} \frac{d\mathbb{P}_{\mathbf{v}'}}{d\mathbb{P}_0} (Y, \mathbf{X}) \right] - 1. \quad (\text{A.29})$$

Notice that Lemma A.1 enables us to compute the right-hand side of (A.29) in closed form. For any $\alpha \in [0, 1]$, function $h_{\alpha}(t) = \cosh[\beta^2/2 \cdot (s-t)] + \alpha^2 \sinh[\beta^2/2 \cdot (s-t)]$ is monotone nonincreasing for $t \in \{0, \dots, s\}$ and $f(s) = 0$. In addition, for any $\mathbf{v} \in \mathcal{H}(s)$ and any $j \in \{0, \dots, s\}$, we define

$$\mathcal{C}_j(\mathbf{v}) = \{\mathbf{v}' \in \mathcal{H}(s) : |\operatorname{supp}(\mathbf{v}) \cap \operatorname{supp}(\mathbf{v}')| = s - j\}. \quad (\text{A.30})$$

For $\ell \in \{1, 2\}$, any query function $q \in \mathcal{Q}_{\mathcal{A}}$, and any $\mathbf{v} \in \mathcal{C}_{\ell}(q)$, by Lemma A.1 and the definition of $\bar{\mathcal{C}}_{\ell}(q, \mathbf{v})$ in (A.28), there exists an integer $k_{\ell}(q, \mathbf{v})$ that satisfies

$$\bar{\mathcal{C}}_{\ell}(q, \mathbf{v}) = \mathcal{C}_0(\mathbf{v}) \cup \mathcal{C}_1(\mathbf{v}) \cup \dots \cup \mathcal{C}_{k_{\ell}(q, \mathbf{v})-1}(\mathbf{v}) \cup \mathcal{C}'_{\ell}(q, \mathbf{v}), \quad (\text{A.31})$$

where $\mathcal{C}'_{\ell}(q, \mathbf{v}) = \bar{\mathcal{C}}_{\ell}(q, \mathbf{v}) \setminus \bigcup_{j=0}^{k_{\ell}(q, \mathbf{v})-1} \mathcal{C}_j(\mathbf{v})$ has cardinality

$$|\mathcal{C}'_{\ell}(q, \mathbf{v})| = |\mathcal{C}_{\ell}(q)| - \sum_{j=0}^{k_{\ell}(q, \mathbf{v})-1} |\mathcal{C}_j(\mathbf{v})| < |\mathcal{C}_{k_{\ell}(q, \mathbf{v})}(\mathbf{v})|. \quad (\text{A.32})$$

Thus we can sandwich the cardinality of $\bar{\mathcal{C}}_{\ell}(q, \mathbf{v})$ by

$$\sum_{j=0}^{k_{\ell}(q, \mathbf{v})} |\mathcal{C}_j(\mathbf{v})| > |\bar{\mathcal{C}}_{\ell}(q, \mathbf{v})| \geq \sum_{j=0}^{k_{\ell}(q, \mathbf{v})-1} |\mathcal{C}_j(\mathbf{v})|. \quad (\text{A.33})$$

Combining Lemmas A.1 and A.4, we further have

$$1 + D_{\chi^2}(\mathbb{P}_{\mathcal{C}_{\ell}(q)}, \mathbb{P}_0) \leq \frac{\sum_{i=0}^{k_{\ell}(q, \mathbf{v})-1} h_{\alpha}(i) \cdot |\mathcal{C}_i(\mathbf{v})| + h_{\alpha}[k_{\ell}(q, \mathbf{v})] \cdot |\mathcal{C}'_{\ell}(q, \mathbf{v})|}{\sum_{j=0}^{k_{\ell}(q, \mathbf{v})-1} |\mathcal{C}_j(\mathbf{v})| + |\mathcal{C}'_{\ell}(q, \mathbf{v})|}, \quad \text{for all } \mathbf{v} \in \mathcal{C}_{\ell}(q). \quad (\text{A.34})$$

Moreover, by (A.34) and the monotonicity of $h_{\alpha}(t)$ we obtain

$$1 + D_{\chi^2}(\mathbb{P}_{\mathcal{C}_{\ell}(q)}, \mathbb{P}_0) \leq \frac{\sum_{i=0}^{k_{\ell}(q, \mathbf{v})-1} h_{\alpha}(i) \cdot |\mathcal{C}_i(\mathbf{v})|}{\sum_{j=0}^{k_{\ell}(q, \mathbf{v})-1} |\mathcal{C}_j(\mathbf{v})|}. \quad (\text{A.35})$$

By the definition of $\mathcal{C}_j(\mathbf{v})$ in (A.30), the cardinality of $\mathcal{C}_j(\mathbf{v})$ does not depend on the choice of $\mathbf{v} \in \mathcal{H}(s)$ and we have $|\mathcal{C}_j(\mathbf{v})| = \binom{s}{s-j} \binom{d-s}{j}$. Thus for any $j \in \{0, \dots, s-1\}$ we have

$$|\mathcal{C}_{j+1}(\mathbf{v})|/|\mathcal{C}_j(\mathbf{v})| = (s-j) \cdot (d-s-j)/(j+1)^2 \geq (d-2s)/s^2. \quad (\text{A.36})$$

Under the assumption that $s^2/d = o(1)$, the right-hand side of (A.36) is lower bounded by $\zeta = d/(2s^2)$ when d and s are sufficiently large. Then we have $|\mathcal{C}_j(\mathbf{v})| \leq \zeta^{j-s} |\mathcal{C}_s(\mathbf{v})|$ for $j \in \{0, \dots, s\}$. By the definition of $k_\ell(q, \mathbf{v})$ in (A.31) and (A.32), for any $q \in \mathcal{Q}_{\mathcal{A}}$, we further obtain

$$\begin{aligned} |\mathcal{C}_\ell(q)| &\leq \sum_{j=0}^{k_\ell(q, \mathbf{v})} |\mathcal{C}_j(\mathbf{v})| \leq |\mathcal{C}_s(\mathbf{v})| \sum_{j=0}^{k_\ell(q, \mathbf{v})} \zeta^{j-s} \\ &\leq \frac{\zeta^{-[s-k_\ell(q, \mathbf{v})]} |\mathcal{H}(s)|}{1 - \zeta^{-1}} \leq 2\zeta^{-[s-k_\ell(q, \mathbf{v})]} |\mathcal{H}(s)|, \end{aligned} \quad (\text{A.37})$$

where the last inequality follows from the fact that $\zeta^{-1} = 2s^2/d = o(1)$.

Moreover, for any two positive sequences $\{a_i\}_{i=0}^s$ and $\{b_i\}_{i=0}^s$ satisfying $a_i/a_{i-1} \geq b_i/b_{i-1} > 1$ for all $i \in [s]$, since $h_\alpha(t)$ is nonincreasing, for any $k \in [s]$, we have

$$\sum_{0 \leq i < j \leq k} (a_i b_j - a_j b_i) \cdot [h_\alpha(i) - h_\alpha(j)] \leq 0. \quad (\text{A.38})$$

Further simplifying the terms in (A.38), we have

$$\frac{\sum_{i=0}^k [a_i h_\alpha(i)]}{\sum_{i=0}^k a_i} \leq \frac{\sum_{i=0}^k [b_i h_\alpha(i)]}{\sum_{i=0}^k b_i}. \quad (\text{A.39})$$

In what follows, we upper bound $k_\ell(q, \mathbf{v})$ for $\ell \in \{1, 2\}$ and $\mathbf{v} \in \mathcal{C}_\ell(q)$. We employ the shorthand $k_\ell = k_\ell(q, \mathbf{v})$ to simplify the notations. Combining (A.29), (A.35), and (A.39) with $a_j = |\mathcal{C}_j(\mathbf{v})|$ and $b_j = \zeta^j$, we have

$$\begin{aligned} 1 + D_{\chi^2}(\mathbb{P}_{\mathcal{C}_\ell(q)}, \mathbb{P}_0) &\leq \frac{\sum_{j=0}^{k_\ell-1} \zeta^j h_\alpha(j)}{\sum_{j=0}^{k_\ell-1} \zeta^j} \\ &= \frac{\sum_{j=0}^{k_\ell-1} \zeta^j \{ \cosh[\beta^2/2 \cdot (s-j)] + \alpha^2 \sinh[\beta^2/2 \cdot (s-j)] \}}{\sum_{j=0}^{k_\ell-1} \zeta^j} \\ &\leq \left\{ \frac{\sum_{j=0}^{k_\ell-1} \zeta^j \cosh[\beta^2(s-j)]}{\sum_{j=0}^{k_\ell-1} \zeta^j} \right\} \vee \left\{ \frac{\sum_{j=0}^{k_\ell-1} \zeta^j \exp[\alpha^2 \beta^2(s-j)]}{\sum_{j=0}^{k_\ell-1} \zeta^j} \right\}. \end{aligned} \quad (\text{A.40})$$

Here the second inequality follows from Lemma A.2. We bound the two terms in (A.40) separately. Note that for notational simplicity, we denote for any $t \in \{0, \dots, s\}$, we define

$$f(t) = \cosh[\beta^2(s-t)], \quad g(t) = \exp[\alpha^2 \beta^2(s-t)].$$

Note that both $h(t)$ and $g(t)$ are monotone non-increasing, and thus $f(t) \geq f(s) = 1$ and $g(t) \geq g(s) = 1$. Moreover, by calculation, we have $f(j-1)/f(j) \geq \cosh(\beta^2)$ for all $j \in \{1, \dots, s\}$. Thus we have

$$f(j) \leq f(k_\ell - 1) \cdot [\cosh(\beta^2)]^{k_\ell - j - 1}, \quad \text{for all } j \in \{0, \dots, k_\ell - 1\}.$$

Then we have

$$\begin{aligned} \frac{\sum_{j=0}^{k_\ell-1} \zeta^j f(j)}{\sum_{j=0}^{k_\ell-1} \zeta^j} &\leq f(k_\ell - 1) \cdot \frac{\sum_{j=0}^{k_\ell-1} \zeta^j [\cosh(\beta^2)]^{k_\ell - j - 1}}{\sum_{j=0}^{k_\ell-1} \zeta^j} \\ &\leq f(k_\ell - 1) \cdot \frac{\sum_{j=0}^{k_\ell-1} [\cosh(\beta^2)/\zeta]^{k_\ell - j - 1}}{\sum_{j=0}^{k_\ell-1} \zeta^{-(k_\ell - j - 1)}} \\ &= f(k_\ell - 1) \cdot \frac{1 - [\cosh(\beta^2)/\zeta]^{k_\ell}}{1 - \zeta^{-k_\ell}} \cdot \frac{1 - \zeta^{-1}}{1 - \zeta^{-1} \cosh(\beta^2)}. \end{aligned} \quad (\text{A.41})$$

Since $\cosh(\beta^2) > 1$, by (A.41) we have

$$\frac{\sum_{j=0}^{k_\ell-1} \zeta^j \cosh[\beta^2(s-j)]}{\sum_{j=0}^{k_\ell-1} \zeta^j} \leq \frac{1 - \zeta^{-1}}{1 - \zeta^{-1} \cosh(\beta^2)} \cdot \cosh[\beta^2(s - k_\ell + 1)]. \quad (\text{A.42})$$

In addition, since $g(j-1)/g(j) = \exp(\alpha^2\beta^2)$, similar to (A.41) we have

$$\frac{\sum_{j=0}^{k_\ell-1} \zeta^j \exp[\alpha^2\beta^2(s-j)]}{\sum_{j=0}^{k_\ell-1} \zeta^j} \leq \frac{1 - \zeta^{-1}}{1 - \zeta^{-1} \exp(\alpha^2\beta^2)} \cdot \exp[\alpha^2\beta^2(s - k_\ell + 1)] \quad (\text{A.43})$$

Combining (A.42) and (A.43), we obtain that

$$1 + D_{\chi^2}(\mathbb{P}_{\mathcal{C}_\ell(q)}, \mathbb{P}_0) \leq \left\{ \frac{(1 - \zeta^{-1}) \cosh[\beta^2(s - k_\ell + 1)]}{1 - \zeta^{-1} \cosh(\beta^2)} \right\} \vee \left\{ \frac{(1 - \zeta^{-1}) \exp[\alpha^2\beta^2(s - k_\ell + 1)]}{1 - \zeta^{-1} \exp(\alpha^2\beta^2)} \right\}. \quad (\text{A.44})$$

Moreover, we use the following lemma obtained from [27] to establish a lower bound for $D_{\chi^2}(\mathbb{P}_{\mathcal{C}_\ell(q)}, \mathbb{P}_0)$.

Lemma A.5. *For any query function q and $\ell \in \{1, 2\}$, we have*

$$D_{\chi^2}(\mathbb{P}_{\mathcal{C}_\ell(q)}, \mathbb{P}_0) \geq \log(T/\xi)/n.$$

We denote $\sqrt{\log(T/\xi)/n}$ by τ for simplicity of notations. Combining (A.44), Lemma A.5 and inequality $\cosh(x) \leq \exp(x^2/2)$, at least one of the two inequality holds

$$(1 + \tau^2) \cdot [1 - \zeta^{-1} \cosh(\beta^2/2)] / (1 - \zeta^{-1}) \leq \exp[\beta^4/2 \cdot (s - k_\ell + 1)^2], \quad (\text{A.45})$$

$$(1 + \tau^2) \cdot [1 - \zeta^{-1} \exp(\alpha^2\beta^2)] / (1 - \zeta^{-1}) \leq \exp[\alpha^2\beta^2(s - k_\ell + 1)]. \quad (\text{A.46})$$

If (A.45) holds, taking the logarithm of the both sides, we have

$$\beta^4/2 \cdot (s - k_\ell + 1)^2 \geq \log(1 + \tau^2) - \log \left[\frac{1 - \zeta^{-1}}{1 - \zeta^{-1} \cosh(\beta^2)} \right]. \quad (\text{A.47})$$

Whereas if (A.46) is true, it holds that

$$\alpha^2\beta^2(s - k_\ell + 1) \geq \log(1 + \tau^2) - \log \left[\frac{1 - \zeta^{-1}}{1 - \zeta^{-1} \exp(\alpha^2\beta^2)} \right]. \quad (\text{A.48})$$

In addition, by Taylor expansion and the fact that $[\cosh(\beta^2/2) + \exp(\alpha^2\beta^2)]/\zeta = o(1)$, we have

$$\log \left[\frac{1 - \zeta^{-1}}{1 - \zeta^{-1} \cosh(\beta^2)} \right] = \log \left\{ 1 + \frac{\zeta^{-1} [\cosh(\beta^2) - 1]}{1 - \zeta^{-1} \cosh(\beta^2)} \right\} = O(\zeta^{-1}\beta^4), \quad (\text{A.49})$$

$$\log \left[\frac{1 - \zeta^{-1}}{1 - \zeta^{-1} \exp(\alpha^2\beta^2)} \right] = \log \left\{ 1 + \frac{\zeta^{-1} [\exp(\alpha^2\beta^2) - 1]}{1 - \zeta^{-1} \exp(\alpha^2\beta^2)} \right\} = O(\zeta^{-1}\alpha^2\beta^2). \quad (\text{A.50})$$

Since $\gamma_n = s\beta^2$, by (A.26) we have $(\alpha^2\beta^2) \vee \beta^4 = o(\log d/n)$. Hence, by (A.49) and (A.50), the second terms on the right-hand sides of (A.47) and (A.48) are asymptotically negligible compared with $\log(1 + \tau^2)$. Therefore, by (A.47) and (A.48), for $\ell \in \{1, 2\}$, at least one of the following two arguments hold:

$$k_\ell(q, \mathbf{v}) \leq s + 1 - \sqrt{\log(1 + \tau^2)/\beta^4}, \quad k_\ell(q, \mathbf{v}) \leq s + 1 - \log(1 + \tau^2)/(2\alpha^2\beta^2).$$

Equivalently, we have

$$k_\ell(q, \mathbf{v}) \leq \left[s + 1 - \sqrt{\log(1 + \tau^2)/\beta^4} \right] \vee \left[s + 1 - \log(1 + \tau^2)/(2\alpha^2\beta^2) \right]. \quad (\text{A.51})$$

Recall that $\tau = \sqrt{\log(T/\xi)/n}$ where $\xi = o(1)$. For any constant $\eta > 0$, we set $T = O(d^\eta)$. By combining Lemmas A.3 and A.4, (A.37), and (A.51), we further obtain

$$T \cdot \frac{\sup_{q \in \mathcal{Q}_{\neq}} (|\mathcal{C}_1(q)| + |\mathcal{C}_2(q)|)}{|\mathcal{H}(s)|} \leq 4T \cdot \exp \left\{ -\log \zeta \cdot \left[\sqrt{\log(1 + \tau^2)/\beta^4} - 1 \right] \right\} \wedge 4T \cdot \exp \left\{ -\log \zeta \cdot \left[\log(1 + \tau^2)/(2\alpha^2\beta^2) - 1 \right] \right\}. \quad (\text{A.52})$$

Under the assumption of the theorem, there is a sufficiently small constant $\delta > 0$ such that $s^2/d^{1-\delta} = O(1)$. Thus we have $\zeta = d/(2s^2) = \Omega(d^\delta)$. By inequality $\log(1 + x) \geq x/2$, it holds that $\log(1 +$

$\tau^2 \geq \tau^2/2 = \log(T/\xi)/(2n)$. Under the condition in (A.26), we have

$$\frac{\log(T/\xi)}{2n\beta^4} \sqrt{\frac{\log(T/\xi)}{4n\alpha^2\beta^2}} > \frac{\log(1/\xi)}{2n\beta^4} \sqrt{\frac{\log(1/\xi)}{4\alpha^2\beta^2}} \rightarrow \infty. \quad (\text{A.53})$$

Hence if n is sufficiently large, the left-hand side in (A.53) is greater than an absolute constant C satisfying $\delta(C-1) > \eta$. Then by (A.52) we have

$$T \cdot \frac{\sup_{q \in \mathcal{Q}_{\mathcal{A}}} (|\mathcal{C}_1(q)| + |\mathcal{C}_2(q)|)}{|\mathcal{H}(s)|} = O[4d^n \zeta^{-(C-1)}] = O[4d^n d^{-\delta(C-1)}] = o(1). \quad (\text{A.54})$$

Combining (A.54) and Lemma A.3, we conclude that $\bar{R}_n^*(\mathcal{G}_0, \mathcal{G}_1; \mathcal{A}, r) \rightarrow 1$ if (A.26) holds. This concludes the proof of Theorem 3.3.

A.4 Proof of Theorem 3.4

To ease notation, we denote the joint distribution of (Y, \mathbf{X}) by $\mathbb{P}_{\boldsymbol{\theta}}$ where the model parameter is given by $\boldsymbol{\theta} = (\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}, \alpha)$. In addition, we let $\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$. Thus $\Delta\boldsymbol{\mu} = \mathbf{0}$ for all $\boldsymbol{\theta} \in \mathcal{G}_0(\boldsymbol{\Sigma})$ and $\Delta\boldsymbol{\mu} \in \mathcal{B}(s)$ for all $\boldsymbol{\theta} \in \mathcal{G}_1(\boldsymbol{\Sigma}; \gamma_n)$. In what follows, we bound the type-I and type-II errors of $\bar{\phi}$ respectively.

Type-I error. For any $\boldsymbol{\theta} \in \mathcal{G}_0(\boldsymbol{\Sigma})$, by the definition of $\bar{\phi}$, the type-I error is bounded by

$$\bar{\mathbb{P}}_{\boldsymbol{\theta}}(\bar{\phi} = 1) \leq \bar{\mathbb{P}}_{\boldsymbol{\theta}}(\bar{\phi}_1 = 1) + \bar{\mathbb{P}}_{\boldsymbol{\theta}}(\bar{\phi}_2 = 1).$$

For test function $\bar{\phi}_1$, since marginally, $\mathbf{X} \sim 1/2 \cdot \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) + 1/2 \cdot \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, for any $\boldsymbol{\theta} \in \mathcal{G}_0(\boldsymbol{\Sigma}) \cup \mathcal{G}_1(\boldsymbol{\Sigma}; \gamma_n)$, for any $j \in [d]$, we have

$$\begin{aligned} & \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}}(X_j^2/\sigma_j - 1) - [\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}}(X_j/\sqrt{\sigma_j})]^2 \\ &= 1/4 \cdot (\mu_{0,j} - \mu_{1,j})^2/\sigma_j = 1/4 \cdot (\Delta\boldsymbol{\mu})_j^2/\sigma_j, \end{aligned} \quad (\text{A.55})$$

Here $\mu_{0,j}$ and $\mu_{1,j}$ denote the j -th entries of $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$, and $(\Delta\boldsymbol{\mu})_j$ is the j -th entry of $\Delta\boldsymbol{\mu}$. In addition, by the definition of q_j in (3.13) we have

$$\begin{aligned} & \left| [\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} q_j(Y, \mathbf{X})]^2 - [\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}}(X_j/\sqrt{\sigma_j})]^2 \right| \\ & \leq 2 |\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}}(X_j/\sqrt{\sigma_j})| \cdot |\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}}(X_j/\sqrt{\sigma_j}) - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} q_j(Y, \mathbf{X})| + |\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}}(X_j/\sqrt{\sigma_j}) - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} q_j(Y, \mathbf{X})|^2. \end{aligned}$$

Since $X_j/\sqrt{\sigma_j} - q_j(Y, \mathbf{X}) = X_j/\sqrt{\sigma_j} \cdot \mathbf{1}\{|X_j/\sqrt{\sigma_j}| > R \cdot \sqrt{\log d}\}$, by Cauchy-Schwarz inequality we have

$$|\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}}(X_j/\sqrt{\sigma_j}) - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} q_j(Y, \mathbf{X})|^2 \leq \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}}(X_j^2/\sigma_j) \cdot \mathbb{P}_{\boldsymbol{\theta}}(|X_j/\sqrt{\sigma_j}| > R \cdot \sqrt{\log d}). \quad (\text{A.56})$$

Since $\|\boldsymbol{\mu}_0\|_{\infty} \vee \|\boldsymbol{\mu}_1\|_{\infty} \leq C_0$ and $\{X_j/\sqrt{\sigma_j}\}_{j=1}^d$ are sub-Gaussian random variables, for any $t > 0$, there exists a constant C_1 such that

$$\mathbb{P}_{\boldsymbol{\theta}}(|X_j/\sqrt{\sigma_j}| > t) \leq 2 \exp(-C_1 t^2). \quad (\text{A.57})$$

Thus setting $t = R \cdot \sqrt{\log d}$ for some sufficiently large R , by (A.56) and (A.57) we obtain

$$|\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}}(X_j/\sqrt{\sigma_j}) - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} q_j(Y, \mathbf{X})| \leq C_2 d^{-1}$$

for some constant C_2 . Thus we have

$$\left| [\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} q_j(Y, \mathbf{X})]^2 - [\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}}(X_j/\sqrt{\sigma_j})]^2 \right| \leq 2C_0 \cdot C_2 d^{-1} + C_2^2 d^{-2} \leq 1/16 \cdot (\Delta\boldsymbol{\mu})_j^2/\sigma_j. \quad (\text{A.58})$$

In addition, since $X_j^2/\sigma_j - 1 - \tilde{q}_j(Y, \mathbf{X}) = (X_j^2/\sigma_j - 1) \cdot \mathbf{1}\{|X_j/\sqrt{\sigma_j}| > R \cdot \sqrt{\log d}\}$, for \tilde{q}_j defined in (3.14), we similarly we obtain

$$|\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} \tilde{q}_j(Y, \mathbf{X}) - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}}(X_j^2/\sigma_j - 1)| \leq 1/16 \cdot (\Delta\boldsymbol{\mu})_j^2/\sigma_j. \quad (\text{A.59})$$

Combining (A.58) and (A.59) we have

$$\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} \tilde{q}_j(Y, \mathbf{X}) - [\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} q_j(Y, \mathbf{X})]^2 \geq 1/8 \cdot (\Delta\boldsymbol{\mu})_j^2/\sigma_j \text{ for all } j \in [d].$$

Taking supremum over $j \in [d]$, we have

$$\sup_{j \in [d]} \left\{ \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} \tilde{q}_j(Y, \mathbf{X}) - [\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} q_j(Y, \mathbf{X})]^2 \right\} \geq 1/8 \cdot \sup_{j \in [d]} [(\Delta\boldsymbol{\mu})_j^2/\sigma_j]. \quad (\text{A.60})$$

Note that the test function $\bar{\phi}$ involves $4d$ queries functions. Thus, for any $\boldsymbol{\theta} \in \mathcal{G}_0(\boldsymbol{\Sigma}) \cup \mathcal{G}_1(\boldsymbol{\Sigma}; \gamma_n)$, under $\mathbb{P}_{\boldsymbol{\theta}}$ the tolerance parameters for q_j and \tilde{q}_j are given by

$$\tau_{q_j} \leq R\sqrt{\log d} \cdot \sqrt{[\log(4d/\xi)]/n}, \quad \tau_{\tilde{q}_j} \leq R^2 \log d \cdot \sqrt{[\log(4d/\xi)]/n}, \quad \text{for all } j \in [d]. \quad (\text{A.61})$$

Under the assumption that

$$\sup_{j \in [d]} (\Delta \boldsymbol{\mu})_j^2 / \sigma_j = \Omega \left[\log^2 d \cdot \log(d/\xi) / (\alpha^2 n) \wedge \log d \cdot \sqrt{\log(d/\xi)/n} \right],$$

we have

$$\begin{aligned} \tau_{q_j} \vee \tau_{\tilde{q}_j} &\leq R^2 \log d \cdot \sqrt{[\log(4d/\xi)]/n} \\ &\leq (1/C) \cdot \left\{ \sup_{j \in [d]} [(\Delta \boldsymbol{\mu})_j^2 / \sigma_j] \vee \alpha \cdot \sup_{j \in [d]} |(\Delta \boldsymbol{\mu})_j / \sqrt{\sigma_j}| \right\}, \end{aligned} \quad (\text{A.62})$$

where the absolute constant C is the same as in (3.16). Note that we denote $R^2 \log d \cdot \sqrt{[\log(4d/\xi)]/n}$ by $\bar{\tau}_1$. Hence by (A.62), for any $\boldsymbol{\theta} \in \mathcal{G}_0(\boldsymbol{\Sigma})$, the type-I error of $\bar{\phi}_1$ is bounded by

$$\begin{aligned} &\bar{\mathbb{P}}_{\boldsymbol{\theta}} \left[\sup_{j \in [d]} (Z_{\tilde{q}_j} - Z_{q_j}^2) \geq C\bar{\tau}_1 \right] \\ &= \bar{\mathbb{P}}_{\boldsymbol{\theta}} \left(\bigcup_{j \in [d]} \left\{ Z_{\tilde{q}_j} - Z_{q_j}^2 - \left\{ \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} \tilde{q}_j(Y, \mathbf{X}) - [\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} q_j(Y, \mathbf{X})]^2 \right\} \geq C\bar{\tau}_1 \right\} \right) \\ &\leq \bar{\mathbb{P}}_{\boldsymbol{\theta}} \left(\bigcup_{j \in [d]} \left\{ Z_{\tilde{q}_j} - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} \tilde{q}_j(Y, \mathbf{X}) \geq \bar{\tau}_1 \right\} \right) + \bar{\mathbb{P}}_{\boldsymbol{\theta}} \left(\bigcup_{j \in [d]} \left\{ Z_{q_j}^2 - [\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} q_j(Y, \mathbf{X})]^2 \geq (C-1)\bar{\tau}_1 \right\} \right). \end{aligned}$$

For the first term, we have

$$\begin{aligned} &\bar{\mathbb{P}}_{\boldsymbol{\theta}} \left(\bigcup_{j \in [d]} \left\{ Z_{\tilde{q}_j} - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} \tilde{q}_j(Y, \mathbf{X}) \geq \bar{\tau}_1 \right\} \right) \\ &\leq \bar{\mathbb{P}}_{\boldsymbol{\theta}} \left(\bigcup_{j \in [d]} \left\{ |Z_{\tilde{q}_j} - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} \tilde{q}_j(Y, \mathbf{X})| \geq \tau_{\tilde{q}_j} \right\} \right) \leq \xi. \end{aligned} \quad (\text{A.63})$$

Note that under the null hypothesis $\boldsymbol{\theta} \in \mathcal{G}_0(\boldsymbol{\Sigma})$, we have $\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} q_j(Y, \mathbf{X}) = \mu_{0,j} / \sqrt{\sigma_j}$. Under the assumption that $\|\boldsymbol{\mu}_0\|_{\infty} \vee \|\boldsymbol{\mu}_1\|_{\infty} \leq C_0$, when n is sufficiently large such that

$$\bar{\tau}_1 \leq 3(C-1)^{-1} C_0 / \sqrt{\sigma_j},$$

by $Z_{q_j}^2 - [\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} q_j(Y, \mathbf{X})]^2 \geq (C-1)\bar{\tau}_1$ we have

$$|Z_{q_j} - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} q_j(Y, \mathbf{X})| \geq (C-1)\bar{\tau}_1 \cdot \sqrt{\sigma_j} / (3C_0). \quad (\text{A.64})$$

Thus we can set absolute constant C sufficiently large such that $|Z_{q_j} - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} q_j(Y, \mathbf{X})| \geq \tau_{q_j}$. Thus by (A.64) we have

$$\begin{aligned} &\bar{\mathbb{P}}_{\boldsymbol{\theta}} \left(\bigcup_{j \in [d]} \left\{ Z_{q_j}^2 - [\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} q_j(Y, \mathbf{X})]^2 \geq (C-1)\bar{\tau}_1 \right\} \right) \\ &\leq \bar{\mathbb{P}}_{\boldsymbol{\theta}} \left(\bigcup_{j \in [d]} \left\{ |Z_{q_j} - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} q_j(Y, \mathbf{X})| \geq \tau_{q_j} \right\} \right) \leq \xi. \end{aligned} \quad (\text{A.65})$$

Combining (A.63) and (A.65), we can bound the type-I error of $\bar{\phi}_1$ by 2ξ . For the type-I error of $\bar{\phi}_2$, we define $\mathbf{Z} = (2Y - 1) \cdot \mathbf{X}$. Under the data-generating model defined in (2.1) and (2.2), the distribution of \mathbf{Z} is given by

$$\mathbf{Z} \sim \frac{1+\alpha}{4} \mathcal{N}(-\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) + \frac{1+\alpha}{4} \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + \frac{1-\alpha}{4} \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) + \frac{1-\alpha}{4} \mathcal{N}(-\boldsymbol{\mu}_1, \boldsymbol{\Sigma}).$$

Then by definition, for all $\boldsymbol{\theta} \in \mathcal{G}_0(\boldsymbol{\Sigma})$, we have

$$\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} [\mathbf{v}^{\top} \text{diag}(\boldsymbol{\Sigma})^{-1/2} \mathbf{Z}] = 0, \quad \text{for all } \mathbf{v} \in \mathcal{B}_2(1). \quad (\text{A.66})$$

In addition, for any $\boldsymbol{\theta} \in \mathcal{G}_1(\boldsymbol{\Sigma}; \gamma_n)$, by the distribution of \mathbf{Z} , for all $\mathbf{v} \in \mathcal{B}_2(1)$, we have

$$\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} [\mathbf{v}^{\top} \text{diag}(\boldsymbol{\Sigma})^{-1/2} \mathbf{Z}] = \alpha/2 \cdot \mathbf{v}^{\top} \text{diag}(\boldsymbol{\Sigma})^{-1/2} \Delta \boldsymbol{\mu}. \quad (\text{A.67})$$

Moreover, by definition we have

$$\mathbf{v}^{\top} \text{diag}(\boldsymbol{\Sigma})^{-1/2} \mathbf{Z} - \bar{q}_{\mathbf{v}}(Y, \mathbf{X}) = \mathbf{v}^{\top} \text{diag}(\boldsymbol{\Sigma})^{-1/2} \mathbf{Z} \cdot \mathbb{1} \left\{ |\mathbf{v}^{\top} \text{diag}(\boldsymbol{\Sigma})^{-1/2} \mathbf{Z}| \leq R\sqrt{\log d} \right\}.$$

By setting the constant R sufficiently large, for any for any $\boldsymbol{\theta} \in \mathcal{G}_0(\boldsymbol{\Sigma}) \cup \mathcal{G}_1(\boldsymbol{\Sigma}; \gamma_n)$, we have

$$|\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} \bar{q}_{\mathbf{v}}(Y, \mathbf{X}) - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} (\mathbf{v}^{\top} \text{diag}(\boldsymbol{\Sigma})^{-1/2} \mathbf{Z})| \leq \alpha/4 \cdot |(\Delta \boldsymbol{\mu})_j / \sqrt{\sigma_j}|.$$

Combining (A.66) and (A.67) we obtain that

$$\begin{aligned}\mathbb{E}_{\mathbb{P}_\theta} \bar{q}_v(Y, \mathbf{X}) &\leq \alpha/4 \cdot |(\Delta\boldsymbol{\mu})_j/\sqrt{\sigma_j}| \text{ for all } \boldsymbol{\theta} \in \mathcal{G}_0(\boldsymbol{\Sigma}); \\ \mathbb{E}_{\mathbb{P}_\theta} \bar{q}_v(Y, \mathbf{X}) &\geq \alpha/4 \cdot |(\Delta\boldsymbol{\mu})_j/\sqrt{\sigma_j}| \text{ for all } \boldsymbol{\theta} \in \mathcal{G}_1(\boldsymbol{\Sigma}, \gamma_n).\end{aligned}$$

Thus, taking the supremum over $\mathcal{B}_2(1)$ yields

$$\sup_{\mathbf{v} \in \mathcal{B}_2(1)} \mathbb{E}_{\mathbb{P}_\theta} \bar{q}_v(Y, \mathbf{X}) \geq \alpha/4 \cdot \sup_{j \in [d]} |(\Delta\boldsymbol{\mu})_j/\sqrt{\sigma_j}| \text{ for all } \boldsymbol{\theta} \in \mathcal{G}_1(\boldsymbol{\Sigma}, \gamma_n). \quad (\text{A.68})$$

In addition, since we have $4d$ queries, by Definition 2.2, the tolerance parameters for \bar{q}_v 's are bounded by

$$\tau_{\bar{q}_v} \leq R\sqrt{\log d} \cdot \sqrt{[\log(4d/\xi)]/n}, \text{ for all } \mathbf{v} \in \mathcal{B}_2(1).$$

Note that we denote $\bar{\tau}_2 = R\sqrt{\log d} \cdot \sqrt{[\log(4d/\xi)]/n}$. Similar to (A.62), we have

$$\tau_{\bar{q}_v} \leq \bar{\tau}_1 \leq (1/C) \cdot \left\{ \sup_{j \in [d]} [(\Delta\boldsymbol{\mu})_j^2/\sigma_j] \vee \sup_{j \in [d]} \alpha |(\Delta\boldsymbol{\mu})_j/\sqrt{\sigma_j}| \right\}. \quad (\text{A.69})$$

Hence by (A.69), for any $\boldsymbol{\theta} \in \mathcal{G}_0(\boldsymbol{\Sigma})$, the type-I error of $\bar{\phi}_2$ is bounded by

$$\begin{aligned}\bar{\mathbb{P}}_\theta \left(\sup_{\mathbf{v} \in \mathcal{B}_2(1)} Z_{\bar{q}_v} \geq 2\bar{\tau}_1 \right) &= \bar{\mathbb{P}}_\theta \left(\bigcup_{\mathbf{v} \in \mathcal{B}_2(1)} \{Z_{\bar{q}_v} - \mathbb{E}_{\mathbb{P}_\theta}[\bar{q}_v(Y, \mathbf{X})] > \bar{\tau}_1\} \right) \\ &\leq \bar{\mathbb{P}}_\theta \left(\bigcup_{\mathbf{v} \in \mathcal{B}_2(1)} \{|Z_{\bar{q}_v} - \mathbb{E}_{\mathbb{P}_\theta}[\bar{q}_v(Y, \mathbf{X})]| \geq \tau_{\bar{q}_v}\} \right) \leq \xi.\end{aligned} \quad (\text{A.70})$$

Combining (A.63), (A.65), and (A.70), we have

$$\bar{\mathbb{P}}_\theta(\bar{\phi} = 1) \leq 3\xi, \text{ for all } \boldsymbol{\theta} \in \mathcal{G}_0(\boldsymbol{\Sigma}).$$

Type-II error. Now we consider $\boldsymbol{\theta} \in \mathcal{G}_1(\boldsymbol{\Sigma}; \gamma_n)$. Note that $\bar{\phi} = 0$ if $\bar{\phi}_1 = 0$ and $\bar{\phi}_2 = 0$. Thus, for any $\boldsymbol{\theta} \in \mathcal{G}_1(\boldsymbol{\Sigma}; \gamma_n)$, we have

$$\bar{\mathbb{P}}_\theta(\bar{\phi} = 0) = \bar{\mathbb{P}}_\theta(\bar{\phi}_1 = 0 \cap \bar{\phi}_2 = 0) \leq \bar{\mathbb{P}}_\theta(\bar{\phi}_1 = 0) \wedge \bar{\mathbb{P}}_\theta(\bar{\phi}_2 = 0).$$

Recall that we denote $\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$. Similar to the proof of Theorem 3.2, we consider two cases of the condition

$$\sup_{j \in [d]} (\Delta\boldsymbol{\mu})_j^2/\sigma_j = \Omega \left[\log(d/\xi)/(\alpha^2 \cdot n) \wedge \sqrt{\log(d/\xi)/n} \right].$$

Case (i). We show that the type-II error of $\bar{\phi}_1$ is negligible under the assumption that

$$\sup_{j \in [d]} (\Delta\boldsymbol{\mu})_j^2/\sigma_j = \Omega \left[\sqrt{\log(d/\xi)/n} \right].$$

Let $j^* = \operatorname{argmax}_{j \in [d]} (\Delta\boldsymbol{\mu})_j^2/\sigma_j$. Then by (A.62), when we have

$$1 + C\bar{\tau} \leq (\Delta\boldsymbol{\mu})_{j^*}^2/\sigma_{j^*} + 1 - C\bar{\tau} = \mathbb{E}_{\mathbb{P}_\theta} \tilde{q}_{j^*}(Y, \mathbf{X}) - [\mathbb{E}_{\mathbb{P}_\theta} q_{j^*}(Y, \mathbf{X})]^2 - C\bar{\tau}. \quad (\text{A.71})$$

Thus combining (A.58), (A.59), and (A.71), we have

$$\begin{aligned}\bar{\mathbb{P}}_\theta \left[\sup_{j \in [d]} (Z_{q_j} - Z_{\bar{q}_j}^2) < C\bar{\tau}_1 \right] \\ \leq \bar{\mathbb{P}}_\theta \left\{ Z_{\tilde{q}_{j^*}} - Z_{q_{j^*}}^2 < \mathbb{E}_{\mathbb{P}_\theta} \tilde{q}_{j^*}(Y, \mathbf{X}) - [\mathbb{E}_{\mathbb{P}_\theta} q_{j^*}(Y, \mathbf{X})]^2 - C\bar{\tau}_1 \right\} \\ \leq \bar{\mathbb{P}}_\theta [\mathbb{E}_{\mathbb{P}_\theta} \tilde{q}_{j^*}(Y, \mathbf{X}) - Z_{\tilde{q}_{j^*}} > \bar{\tau}_1] + \bar{\mathbb{P}}_\theta \left\{ [\mathbb{E}_{\mathbb{P}_\theta} q_{j^*}(Y, \mathbf{X})]^2 - Z_{q_{j^*}}^2 > (C-1)\bar{\tau}_1 \right\}.\end{aligned} \quad (\text{A.72})$$

Moreover, by (A.62) the first term on the right-hand side of (A.72) can be further bounded by

$$\begin{aligned}\bar{\mathbb{P}}_\theta [\mathbb{E}_{\mathbb{P}_\theta} \tilde{q}_{j^*}(Y, \mathbf{X}) - Z_{\tilde{q}_{j^*}} > \bar{\tau}_1] \\ \leq \bar{\mathbb{P}}_\theta \{ \mathbb{E}_{\mathbb{P}_\theta} \tilde{q}_{j^*}(Y, \mathbf{X}) - Z_{\tilde{q}_{j^*}} \geq \tau_{\tilde{q}_{j^*}} \} \\ \leq \bar{\mathbb{P}}_\theta \left(\bigcup_{j \in [d]} \{|Z_{\tilde{q}_j} - \mathbb{E}_{\mathbb{P}_\theta} \tilde{q}_j(Y, \mathbf{X})| \geq \tau_{\tilde{q}_j}\} \right) \leq \xi.\end{aligned} \quad (\text{A.73})$$

Similarly, for the second term on the right-hand side of (A.72), by (A.62) and (A.64) we have

$$\begin{aligned} & \bar{\mathbb{P}}_{\boldsymbol{\theta}} \left\{ [\mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} q_{j^*}(Y, \mathbf{X})]^2 - Z_{q_{j^*}}^2 > (C-1)\bar{\tau}_1 \right\} \\ & \leq \bar{\mathbb{P}}_{\boldsymbol{\theta}} \left(\bigcup_{j \in [d]} \{ |Z_{q_j} - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} \tilde{q}_j(Y, \mathbf{X})| \geq \tau_{q_j} \} \right) \leq \xi. \end{aligned} \quad (\text{A.74})$$

Therefore, combining (A.73) and (A.74), we conclude that the type-II error of $\bar{\phi}_2$ is no more than 2ξ .

Case (ii). Now we assume study the type-II error of $\bar{\phi}_2$ under the assumption that

$$\sup_{j \in [d]} (\Delta \boldsymbol{\mu})_j^2 / \sigma_j = \Omega [\log(d/\xi) / (\alpha^2 \cdot n)].$$

Let $j^* = \operatorname{argmax}_{j \in [d]} (\Delta \boldsymbol{\mu})_j^2 / \sigma_j$ and $\mathbf{v}^* = \operatorname{argmax}_{\mathbf{v} \in \mathcal{B}_2(1)} \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} \bar{q}_{\mathbf{v}}(Y, \mathbf{X})$. Then by (A.62) and (A.68), when $C > 4$ we have

$$2\bar{\tau}_2 \leq \alpha/2 \cdot \sup_{j \in [d]} |(\Delta \boldsymbol{\mu})_j / \sqrt{\sigma_j}| - 2\bar{\tau}_2 = \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} \bar{q}_{\mathbf{v}^*}(Y, \mathbf{X}) - 2\bar{\tau}_2. \quad (\text{A.75})$$

Then by (A.69) and (A.75) the type-II error of $\bar{\phi}_2$ is bounded by

$$\begin{aligned} & \bar{\mathbb{P}}_{\boldsymbol{\theta}} \left(\sup_{\mathbf{v} \in \mathcal{B}_2(1)} Z_{\bar{q}_{\mathbf{v}}} < 2\bar{\tau}_2 \right) \leq \bar{\mathbb{P}}_{\boldsymbol{\theta}} \left[\sup_{\mathbf{v} \in \mathcal{B}_2(1)} Z_{\bar{q}_{\mathbf{v}}} < \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} \bar{q}_{\mathbf{v}^*}(Y, \mathbf{X}) - 2\bar{\tau}_2 \right] \\ & \leq \bar{\mathbb{P}}_{\boldsymbol{\theta}} [Z_{\bar{q}_{\mathbf{v}^*}} < \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} \bar{q}_{\mathbf{v}^*}(Y, \mathbf{X}) - 2\bar{\tau}_2] \\ & \leq \bar{\mathbb{P}}_{\boldsymbol{\theta}} \left(\bigcup_{\mathbf{v} \in \mathcal{B}_2(1)} \{ |Z_{\bar{q}_{\mathbf{v}}} - \mathbb{E}_{\mathbb{P}_{\boldsymbol{\theta}}} [\bar{q}_{\mathbf{v}}(Y, \mathbf{X})]| \geq \tau_{\bar{q}_{\mathbf{v}}} \} \right) \leq \xi. \end{aligned} \quad (\text{A.76})$$

Thus by (A.76), the type-II error of $\bar{\phi}_2$ is no more than ξ . Then together with Case (i), we have $\bar{\mathbb{P}}_{\boldsymbol{\theta}}(\bar{\phi}) \leq 2\xi$ for all $\boldsymbol{\theta} \in \mathcal{G}_1(\boldsymbol{\Sigma}; \gamma_n)$. Therefore the total risk of $\bar{\phi}$ is bounded by

$$\bar{R}_n(\bar{\phi}) = \sup_{\boldsymbol{\theta} \in \mathcal{G}_0(\boldsymbol{\Sigma})} \bar{\mathbb{P}}_{\boldsymbol{\theta}}(\bar{\phi} = 1) + \sup_{\boldsymbol{\theta} \in \mathcal{G}_1(\boldsymbol{\Sigma}; \gamma_n)} \bar{\mathbb{P}}_{\boldsymbol{\theta}}(\bar{\phi} = 0) \leq 5\xi.$$

B Proofs for Technical Lemmas

In this section, we prove the technical lemmas which appear in the proofs of the main results.

B.1 Proof of Lemma A.1

Under $\mathbb{P}_{\mathbf{0}}$, \mathbf{X} and Y are independent with $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and Y is uniform over $\{0, 1\}$. We denote by $f(\mathbf{x}; \boldsymbol{\mu})$ the density of $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$ and by $p_{\mathbf{0}}(y, \mathbf{x})$ the density of $\mathbb{P}_{\mathbf{0}}$. Then for any $y \in \{0, 1\}$ and $\mathbf{x} \in \mathbb{R}^d$, we have $p_{\mathbf{0}}(y, \mathbf{x}) = 1/2 \cdot f(\mathbf{x}; \mathbf{0})$. In addition, for any $\mathbf{v} \in \mathcal{H}(s)$, we denote the density of $\mathbb{P}_{\mathbf{v}}$ by $p_{\mathbf{v}}(y, \mathbf{x})$. By the definition of the statistical model, we have

$$\begin{aligned} p_{\mathbf{v}}(1, \mathbf{x}) &= (1 + \alpha)/4 \cdot f(\mathbf{x}; \mathbf{v}/2) + (1 - \alpha)/4 \cdot f(\mathbf{x}; -\mathbf{v}/2), \\ p_{\mathbf{v}}(0, \mathbf{x}) &= (1 - \alpha)/4 \cdot f(\mathbf{x}; \mathbf{v}/2) + (1 + \alpha)/4 \cdot f(\mathbf{x}; -\mathbf{v}/2). \end{aligned}$$

Thus for any $y \in \{0, 1\}$ and $\mathbf{x} \in \mathbb{R}^d$, we have

$$\frac{d\mathbb{P}_{\mathbf{v}}}{d\mathbb{P}_{\mathbf{0}}}(y, \mathbf{x}) = \frac{1}{2} \cdot \left[\frac{f(\mathbf{x}; \mathbf{v}/2)}{f(\mathbf{x}; \mathbf{0})} + \frac{f(\mathbf{x}; -\mathbf{v}/2)}{f(\mathbf{x}; \mathbf{0})} \right] + \frac{\alpha(2y-1)}{2} \cdot \left[\frac{f(\mathbf{x}; \mathbf{v}/2)}{f(\mathbf{x}; \mathbf{0})} - \frac{f(\mathbf{x}; -\mathbf{v}/2)}{f(\mathbf{x}; \mathbf{0})} \right]. \quad (\text{B.1})$$

Note that by definition, for any $\boldsymbol{\mu} \in \mathbb{R}^d$, we have

$$g(\mathbf{x}; \boldsymbol{\mu}) := f(\mathbf{x}; \boldsymbol{\mu}) / f(\mathbf{x}; \mathbf{0}) = \exp(\boldsymbol{\mu}^\top \mathbf{x} - 1/2 \cdot \|\boldsymbol{\mu}\|_2^2).$$

Thus (B.1) is reduced to

$$\frac{d\mathbb{P}_{\mathbf{v}}}{d\mathbb{P}_{\mathbf{0}}}(y, \mathbf{x}) = [g(\mathbf{x}, \mathbf{v}/2) + g(\mathbf{x}; -\mathbf{v}/2)] / 2 + \alpha(2y-1) \cdot [g(\mathbf{x}, \mathbf{v}/2) - g(\mathbf{x}; -\mathbf{v}/2)] / 2. \quad (\text{B.2})$$

For any $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{H}(s)$, by (B.2) we have

$$\begin{aligned} & \mathbb{E}_{\mathbb{P}_{\mathbf{0}}} \left[\frac{d\mathbb{P}_{\mathbf{v}_1}}{d\mathbb{P}_{\mathbf{0}}} \frac{d\mathbb{P}_{\mathbf{v}_2}}{d\mathbb{P}_{\mathbf{0}}}(Y, \mathbf{X}) \right] \\ &= \mathbb{E}_{\mathbb{P}_{\mathbf{0}}} \{ [g(\mathbf{X}, \mathbf{v}_1/2) + g(\mathbf{X}; -\mathbf{v}_1/2)] \cdot [g(\mathbf{X}, \mathbf{v}_2/2) + g(\mathbf{X}; -\mathbf{v}_2/2)] / 4 \} \\ & \quad + \alpha^2 \cdot \mathbb{E}_{\mathbb{P}_{\mathbf{0}}} \{ [g(\mathbf{X}, \mathbf{v}_1/2) - g(\mathbf{X}; -\mathbf{v}_1/2)] \cdot [g(\mathbf{X}, \mathbf{v}_2/2) - g(\mathbf{X}; -\mathbf{v}_2/2)] / 4 \}, \end{aligned} \quad (\text{B.3})$$

where we use the independence of Y and \mathbf{X} under \mathbb{P}_0 . In what follows, we calculate the two terms on the right-hand side of (B.3), respectively. Let η_1 and η_2 be two independent Rademacher random variables over $\{-1, 1\}$. Then for $\ell \in \{1, 2\}$, we have

$$[g(\mathbf{X}, \mathbf{v}_\ell/2) + g(\mathbf{X}; -\mathbf{v}_\ell/2)]/2 = \mathbb{E}_{\eta_\ell} [g(\mathbf{X}, \eta_\ell \mathbf{v}_\ell/2)], \quad (\text{B.4})$$

$$[g(\mathbf{X}, \mathbf{v}_\ell/2) - g(\mathbf{X}; -\mathbf{v}_\ell/2)]/2 = \mathbb{E}_{\eta_\ell} [\eta_\ell \cdot g(\mathbf{X}, \eta_\ell \mathbf{v}_\ell/2)]. \quad (\text{B.5})$$

Then by (B.4) and (B.5) we have

$$\begin{aligned} & \mathbb{E}_{\mathbb{P}_0} \{ [g(\mathbf{X}, \mathbf{v}_1/2) + g(\mathbf{X}; -\mathbf{v}_1/2)] \cdot [g(\mathbf{X}, \mathbf{v}_2/2) + g(\mathbf{X}; -\mathbf{v}_2/2)] / 4 \} \\ &= \mathbb{E}_{\mathbb{P}_0} \mathbb{E}_{\eta_1, \eta_2} [g(\mathbf{X}; \eta_1 \mathbf{v}_1/2) \cdot g(\mathbf{X}; \eta_2 \mathbf{v}_2/2)] \\ &= \mathbb{E}_{\eta_1, \eta_2} \mathbb{E}_{\mathbb{P}_0} \exp [\mathbf{X}^\top (\eta_1 \mathbf{v}_1 + \eta_2 \mathbf{v}_2) / 2 - 1/8 \cdot (\|\mathbf{v}_1\|_2^2 + \|\mathbf{v}_2\|_2^2)]. \end{aligned} \quad (\text{B.6})$$

Using the moment-generating function of \mathbf{X} , by (B.6) we have

$$\begin{aligned} & \mathbb{E}_{\mathbb{P}_0} \{ [g(\mathbf{X}, \mathbf{v}_1/2) + g(\mathbf{X}; -\mathbf{v}_1/2)] \cdot [g(\mathbf{X}, \mathbf{v}_2/2) + g(\mathbf{X}; -\mathbf{v}_2/2)] / 4 \} \\ &= \mathbb{E}_{\eta_1, \eta_2} [\exp(1/2 \cdot \eta_1 \eta_2 \cdot \mathbf{v}_1^\top \mathbf{v}_2) = \cosh(1/2 \cdot \langle \mathbf{v}_1, \mathbf{v}_2 \rangle)]. \end{aligned}$$

Similarly, for (B.5) we have

$$\begin{aligned} & \mathbb{E}_{\mathbb{P}_0} \{ [g(\mathbf{X}, \mathbf{v}_1/2) - g(\mathbf{X}; -\mathbf{v}_1/2)] \cdot [g(\mathbf{X}, \mathbf{v}_2/2) - g(\mathbf{X}; -\mathbf{v}_2/2)] / 4 \} \\ &= \mathbb{E}_{\mathbb{P}_0} \mathbb{E}_{\eta_1, \eta_2} [\eta_1 \eta_2 \cdot g(\mathbf{X}; \eta_1 \mathbf{v}_1/2) \cdot g(\mathbf{X}; \eta_2 \mathbf{v}_2/2)] \\ &= \mathbb{E}_{\eta_1, \eta_2} [\eta_1 \eta_2 \cdot \exp(1/2 \cdot \eta_1 \eta_2 \cdot \mathbf{v}_1^\top \mathbf{v}_2)] = \sinh(1/2 \cdot \langle \mathbf{v}_1, \mathbf{v}_2 \rangle). \end{aligned}$$

Thus we conclude the proof of Lemma A.1.

B.2 Proof of Lemma A.2

It is straightforward to verify (A.4) holds when $x = 0$. We focus on region $x > 0$. It is then sufficient to prove the result for these two cases below.

Case 1: We consider the case $v \leq 1/(2x) \cdot \log[\cosh(2x)]$. Then we need to prove

$$\cosh(x) + v \sinh(x) \leq \cosh(2x). \quad (\text{B.7})$$

Using the bound of v , it remains to show the function

$$f(x) = 1/(2x) \cdot \log[\cosh(2x)] \cdot \sinh(x) + \cosh(x) - \cosh(2x) \leq 0.$$

holds for all $x > 0$. It's easy to verify $f(x)$ is monotonically decreasing over $(0, \infty]$ and $\lim_{x \rightarrow 0} f(x) = 0$. We thus finish proving (B.7).

Case 2: We consider the case $v \geq 1/(2x) \cdot \log[\cosh(2x)]$. We would like to show

$$\cosh(x) + v \sinh(x) \leq \exp(2vx). \quad (\text{B.8})$$

Let us define $g(v) := \exp(2vx) - \cosh(x) - v \sinh(x)$. We have that for any $x \geq 0$,

$$g'(v) = 2x \exp(2vx) - \sinh(x) \geq 2x \cosh(2x) - \sinh(x) \geq 0.$$

Hence, $g(v)$ is a monotonically increasing function. We thus have

$$\begin{aligned} g(v) &\geq g\{1/(2x) \cdot \log[\cosh(2x)]\} \\ &= \cosh(2x) - \cosh(x) - 1/(2x) \cdot \log[\cosh(2x)] \cdot \sinh(x) = -f(x) \geq 0. \end{aligned}$$

We thus finish proving (B.8).

C Supporting Lemmas

In this section we list the supporting lemmas that establish two concentration inequalities for Gaussian random variables.

Lemma C.1 (χ^2 -tail bound, [16]). *Let X_1, \dots, X_n be n i.i.d. standard normal random variables. For all $t \in (0, 1)$,*

$$\mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^n X_i^2 - 1 \right| \geq t \right) \leq 2 \exp(-nt^2/8).$$

Lemma C.2 (Gaussian covariance estimation, [25]). *Suppose $\{\mathbf{X}_i\}_{i=1}^n$ are n i.i.d. Gaussian random vectors in \mathbb{R}^d and $\mathbf{X}_1 \sim \mathcal{N}(\mathbf{0}, \Sigma)$. For every $\epsilon \in (0, 1)$, and $t \geq 1$, if $n \geq C(t/\epsilon)^2 d$ for some*

constant C , then with probability at least $1 - 2e^{-t^2 n}$,

$$\|\widehat{\Sigma} - \Sigma\|_2 \leq \epsilon \|\Sigma\|_2,$$

where $\widehat{\Sigma} := 1/n \cdot \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^\top$.