Journal of Machine Learning Research 25 (2024) 1-61 Submitted 7/22; Revised 4/24; Published 7/24

Non-splitting Neyman-Pearson Classifiers

Jingming Wang [*] PDW9QV@VIRGINIA.EDU
Department of Statistics
University of Virginia

Lucy Xia [*] LUCYXIA@UST.HK
Department of ISOM
School of Business and Management
Hong Kong University of Science and Technology

Zhigang Bao ZGBAO@HKU.HK
Department of Mathematics
The University of Hong Kong

Xin Tong XINT@MARSHALL.USC.EDU
Department of Data Sciences and Operations
Marshall Business School
University of Southern California

Editor: Samory Kpotufe

Abstract

The Neyman-Pearson (NP) binary classification paradigm constrains the more severe type of error (e.g., the type I error) under a preferred level while minimizing the other (e.g., the type II error). This paradigm is suitable for applications such as severe disease diagnosis, fraud detection, among others. A series of NP classifiers have been developed to guarantee the type I error control with high probability. However, these existing classifiers involve a sample splitting step: a mixture of class 0 and class 1 observations to construct a scoring function and some left-out class 0 observations to construct a threshold. This splitting enables classifier threshold construction built upon independence, but it amounts to insufficient use of data for training and a potentially higher type II error. Leveraging a canonical linear discriminant analysis (LDA) model, we derive a quantitative CLT for a certain functional of quadratic forms of the inverse of sample and population covariance matrices, and based on this result, develop for the first time NP classifiers without splitting the training sample. Numerical experiments have confirmed the advantages of our new non-splitting parametric strategy.

Keywords: classification, Neyman-Pearson (NP), type I error, non-splitting, efficiency.

[*] Jingming Wang and Lucy Xia contribute equally to the work.

1. Introduction

Classification aims to accurately assign class labels (e.g., fraud vs. non-fraud) to new observations (e.g., new credit card transactions) on the basis of labeled observations (e.g., labeled transactions). The prediction is usually not perfect. In transaction fraud detection, two errors might arise: (1) mislabeling a fraudulent transaction as non-fraudulent and (2) mislabeling a non-fraudulent transaction as fraudulent. The consequences of the two errors are different: while declining a legitimate transaction may cause temporary inconvenience for a consumer, approving a fraudulent transaction can result in a substantial financial loss for a credit card company. In severe disease diagnosis (e.g., cancer vs. normal), the asymmetry of the two errors' importance is even greater: while misidentifying a healthy person as ill may cause anxiety and create additional medical expenses, telling cancer patients that they are healthy may cost their lives. In these applications and beyond, it is critical to prioritize the control of the more important error (Reeve et al., 2023; Müller et al., 2023).

Most theoretical work on binary classification concerns risk. Risk is a weighted sum of type I error (i.e., the conditional probability that the predicted label is 1 given that the true label is 0) and type II error (i.e., the conditional probability that the predicted label is 0 given that the true label is 1), where the weights are marginal probabilities of the two class labels. In the context of transaction fraud detection, coding the fraud class as 0, we would like to control type I error under some small level. The common *classical paradigm*, which minimizes the risk, does not guarantee delivery of classifiers that have type I error bounded by the preferred level. To address this concern, we can employ a general statistical framework for controlling asymmetric errors in binary classification: the *Neyman-Pearson (NP) classification paradigm*, which seeks a classifier that minimizes type II error subject to type I error $\leq \alpha$, where α is a user-specified level, usually a small value (e.g., 5%). The NP framework can achieve the best type II error given a high priority on the type I error.

The NP approach is fundamental in hypothesis testing (justified by the NP lemma), but its use in classification did not occur until the 21st century (Cannon et al., 2002; Scott and Nowak, 2005; Scott, 2007). In the past ten years, there has been significant progress in the theoretical/methodological investigation of NP classification. An incomplete overview includes (i) a theoretical evaluation criterion for NP classifiers: the NP oracle inequalities (Rigollet and Tong, 2011), (ii) classifiers satisfying this criterion under different settings (Tong, 2013; Zhao et al., 2016; Tong et al., 2020), and (iii) practical algorithms for constructing NP classifiers (Tong et al., 2018, 2020), (iv) generalizations to domain adaptation (Scott, 2019) and to multi-class (Tian and Feng, 2021), and (v) theoretical results on minimax rates (Kalan and Kpotufe, 2024).

Unlike the oracle classifier under the classical paradigm, which thresholds the regression function at precisely 1/2, the threshold of the NP oracle is α-dependent and needs to be estimated when we construct sample-based classifiers. Threshold determination is the key in NP classification algorithms, because it is subtle to ensure a high probability control on the type I error under α while achieving satisfactory type II error performance.

For existing NP classification algorithms (Tong, 2013; Zhao et al., 2016; Tong et al., 2018, 2020), a sample splitting step is common practice: a mixture of class 0 and class 1 observations to construct a scoring function $\hat{s}(\cdot)$ (e.g., fitted sigmoid function in logistic regression) and some left-out class 0 observations $\{x_1^0, \cdots, x_m^0\}$ to construct a threshold. Then under proper sampling assumptions, conditioning on $\hat{s}(\cdot)$, the set $\{s_1 := \hat{s}(x_1^0), \cdots, s_m :=$

$\hat{s}(x_m^0)\}$ consists of independent elements. This independence is important in the subsequent threshold determination and classifier construction. Let us take the NP umbrella algorithm introduced in Tong et al. (2018) as an example: it constructs an NP classifier $\hat{\varphi}_\alpha(\cdot) = \mathbb{I}(\hat{s}(\cdot) > s_{(k^*)})$, where $\mathbb{I}(\cdot)$ is the indicator function, $s_{(k^*)}$ is the k^*th order statistic in $\{s_1, \cdots, s_m\}$ and $k^* = \min\left\{k \in \{1, \cdots, m\} : \sum_{j=k}^m \binom{m}{j}(1-\alpha)^j \alpha^{m-j} \leq \delta\right\}$. The smallest order was chosen to have the best type II error. The type I error violation rate has been shown to satisfy $\mathbb{P}(R_0(\hat{\varphi}_\alpha) > \alpha) \leq \sum_{i=k^*}^m \binom{m}{j}(1-\alpha)^j \alpha^{m-j}$, where R_0 denotes the (population-level) type I error. Hence with probability at least $1-\delta$, we have $R_0(\hat{\varphi}_\alpha) \leq \alpha$. Without the independence of $\{s_1, \cdots, s_m\}$, the upper bound on the violation rate does not hold. Therefore, if we used up all class 0 observations in constructing $\hat{s}(\cdot)$, this umbrella algorithm fails. The NP umbrella algorithm tends to be conservative in controlling the type I error and thus may lead to an undesirably large type II error. In other NP works (Tong, 2013; Zhao et al., 2016; Tong et al., 2020), the independence is necessary in threshold determination when applying Vapnik-Chervonenkis inequality, Dvoretzky-Kiefer-Wolfowitz inequality, or constructing classic t-statistics, respectively.

In general, setting aside part of class 0 sample lowers the quality of the scoring function $\hat{s}(\cdot)$, and therefore makes the type II error deteriorate. This becomes a serious concern when the class 0 sample size is small. A more data-efficient alternative is to use all data to construct the scoring function, but this would lose the critical independence property when constructing the threshold. Innovating a non-splitting strategy has long been on the "wish list." This is an important but challenging task. For example, the NP umbrella algorithm, which has no assumption on data distribution and adapts all scoring-type classification methods (e.g., logistic regression, neural nets) to the NP paradigm universally via the nonparametric order statistics approach, has little potential to be extended to the non-splitting scenario, simply because there is no way to characterize the general dependence. To address it, we need to start from tractable distributional assumptions.

Among the commonly used models for classification is the linear discriminant analysis (LDA) model (Hastie et al., 2009; James et al., 2014; Fan et al., 2020), which assumes that the two class-conditional feature distributions are Gaussian with different means but a common covariance matrix: $\mathcal{N}(\boldsymbol{\mu}^0, \Sigma)$ and $\mathcal{N}(\boldsymbol{\mu}^1, \Sigma)$. Classifiers based on the LDA model have been popular in the literature (Shao et al., 2011; Fan et al., 2012; Witten and Tibshirani, 2012; Mai et al., 2012; Hao et al., 2015; Pan et al., 2016; Wang and Jiang, 2018; Cai and Zhang, 2019; Li and Lei, 2018; Sifaou et al., 2020). Hence, it is natural to start our inquiry with the LDA model. However, even this canonical model demands novel intermediate technical results that were not available in the literature. For example, we will need delicate expansion results of quadratic forms of the inverse of sample and population covariance matrices, which we establish for the first time in this manuscript.

As the first effort to investigate a non-splitting strategy under the NP paradigm, this work addresses basic settings. We only work in the regime that $p/n \to [0, 1)$, where p is the feature dimensionality and n is the sample size. We take minimum assumptions on Σ and $\boldsymbol{\mu}_d := \boldsymbol{\mu}^1 - \boldsymbol{\mu}^0$: $\boldsymbol{\mu}_d^\top \Sigma^{-1} \boldsymbol{\mu}_d$ is bounded from below. We do not have specific structural assumptions on Σ or $\boldsymbol{\mu}_d$ such as sparsity. With these minimal assumptions, we propose our new classifier `eLDA` (where `e` stands for *data efficiency*) based on a quantitative CLT for a certain functional of quadratic forms of the inverse of sample and population covariance matrices and show that eLDA respects the type I error control with high probability. Moreover, if $p/n \to 0$, the excess type II error of eLDA, that is the difference between the type II error of eLDA and that of the NP oracle, diminishes as the sample size increases; if $p/n \to r_0 \in (0,1)$, the excess type II error of eLDA diminishes if and only if $\boldsymbol{\mu}_d^\top \Sigma^{-1} \boldsymbol{\mu}_d$ diverges. We note in particular that this work is the first one to establish lower bound results on excess type II error under the NP paradigm.

Table 1: eLDA vs. pNP-LDA vs. NP-sLDA

	eLDA	pNP-LDA	NP-sLDA
type I error	.0314	.0037	.0215
type II error	.4478	.7638	.6102

In addition to enjoying good theoretical properties, eLDA has numerical advantages. Here we take a toy example: $\Sigma = I$, $\boldsymbol{\mu}_d = (1.2, 1.2, 1.2)^\top$ and $\boldsymbol{\mu}^0 = (0,0,0)^\top$. The sample sizes n_0 and n_1 for classes 0 and 1 respectively are both 50. We set the type I error upper bound $\alpha = 0.05$ and the type I error violation rate target $\delta = 0.1$. In this situation, if we were to use the NP umbrella algorithm, we would have to reserve at least 45 (i.e., $\lceil \log \delta / \log(1-\alpha) \rceil$) class 0 observations for threshold determination, and thus at most 5 class 0 observations can be used for scoring function training. With the default setting of a 50-50 split of n_0 in the umbrella algorithm, only 25 observations would be reserved for estimating the threshold, rendering the NP umbrella algorithm not applicable. Even if we skew the split ratio and allocate 90% of n_0 to construct the threshold, the performance of the NP umbrella algorithm would not be desirable. Concretely, we present the results in Table 1, which compares the performance of NP-sLDA (NP umbrella algorithm with sparse LDA scoring function) with two other methods that do not explicitly require a minimum n_0: the newly proposed eLDA, and pNP-LDA, another LDA-based classifier introduced in Tong et al. (2020) that relies on sample splitting and explicit parametric assumptions for threshold determination. In Table 1, the type I and type II errors are averaged over 1,000 repetitions and evaluated on a large test set (50,000 observations from each class) that approximates the population. The results clearly demonstrate that our new non-splitting eLDA classifier outperforms the splitting pNP-LDA classifier by achieving a significantly smaller type II error. This observation is not coincidental. When the more generic nonparametric NP umbrella algorithm is not applicable due to sample size limitations, eLDA typically outperforms pNP-LDA. The results also highlight the advantage of eLDA over NP-sLDA (a skewed 90% − 10% split version) by achieving a much smaller type II error.

This point is further substantiated by an extensive real data analysis conducted in this paper. We consider a wide range of datasets, including medical datasets with small sample sizes, as well as image and cybersecurity datasets commonly studied in the machine learning community. Our findings demonstrate that although eLDA is developed based on a seemingly restrictive LDA model, its performance surpasses that of its existing competitors in small sample regimes. In these medical datasets, it is evident that the NP umbrella algorithm fails to meet the sample size requirement under specified α and δ, while eLDA not only functions effectively but also clearly outperforms pNP-LDA in terms of type II error. In the case of the spam email dataset, where the selected covariates have significantly non-Gaussian distributions, eLDA consistently outperforms its competitors in controlling type I errors, demonstrating the robustness of eLDA beyond the LDA model. In summary, the applicability of `eLDA` extends beyond the Gaussian model and theories on which it is developed.

The rest of the paper is organized as follows. In Section 2, we introduce the essential notations and assumptions. In Section 3, we derive the efficient non-splitting NP classifier `eLDA` and its close relative `feLDA`, where `f` stands for *fixed feature dimension*, and show their main theoretical results. Technical preliminaries are presented in Section 4.1, followed by key technical results in Section 4.2 and the proof of the main theorem in Section 4.3. In Section 5, we present simulation and real data studies. We provide a short discussion in Section 6. In addition, in Appendix A, we make further remarks on our assumptions. The proofs of other theoretical results except for the main theorem are relegated to Appendix B. In Appendix C, we provide the proofs of the technical preliminaries in Section 4.1, followed by the proofs of the key lemmas in the proof of the main theorem in Appendix D. Finally, Appendix E collects additional numerical results.

2. Model and Setups

Let $\phi : \mathcal{X} \subset \mathbb{R}^p \to \{0, 1\}$ denote a mapping from the feature space to the label space. The level-α NP oracle $\phi^*_\alpha(\cdot)$ is defined as the solution to the program $\min_{R_0(\phi)\leq\alpha} R_1(\phi)$, where $R_0(\phi) = \mathbb{P}\{\phi(\mathbf{x}) \neq Y \big| Y = 0\}$ and $R_1(\phi) = \mathbb{P}\{\phi(\mathbf{x}) \neq Y \big| Y = 1\}$ denote the (population-level) type I and type II errors of $\phi(\cdot)$, respectively. We assume the linear discriminant analysis (LDA) model, i.e., $(\mathbf{x}|Y = 0) \sim \mathcal{N}(\boldsymbol{\mu}^0, \Sigma)$ and $(\mathbf{x}|Y = 1) \sim \mathcal{N}(\boldsymbol{\mu}^1, \Sigma)$, where $\boldsymbol{\mu}^0, \boldsymbol{\mu}^1 \in \mathbb{R}^p$ and the common positive definite covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$. Under the LDA model, the level-α NP oracle classifier can be derived explicitly as

$$\phi^*_\alpha(x) = \mathbb{1}\Big((\Sigma^{-1}\boldsymbol{\mu}_d)^\top x > \sqrt{\boldsymbol{\mu}_d^\top \Sigma^{-1}\boldsymbol{\mu}_d}\, \Phi^{-1}(1-\alpha) + \boldsymbol{\mu}_d^\top \Sigma^{-1}\boldsymbol{\mu}^0\Big)\,, \tag{2.1}$$

in which $\boldsymbol{\mu}_d = \boldsymbol{\mu}^1 - \boldsymbol{\mu}^0$, and $\Phi^{-1}(1-\alpha)$ denotes the $(1-\alpha)$-th quantile of standard normal distribution.

For readers' convenience, we introduce a few notations together. For any $k \in \mathbb{N}$, let I_k denote the identity matrix of size k, $\mathbf{1}_k$ denote the all-one column vector of dimension k. For arbitrary two column vectors $\mathbf{u}, \mathbf{v}$ of dimensions a, b, respectively, and any $a \times b$ matrix M, we write $(M)_{\mathbf{uv}}$ as the quadratic form $\mathbf{u}^\top M\mathbf{v}$. Moreover, we write M_{ij} or $(M)_{ij}$ for $i \in \{1, \cdots, a\}$ and $j \in \{1, \cdots, b\}$ as the (i, j)-th entry of M. We use $\|A\|$ to denote the operator norm for a matrix A and use $\|\mathbf{v}\|$ to denote the ℓ_2 norm of a vector $\mathbf{v}$. For two positive sequences A_n and B_n, we adopt the notation $A_n \asymp B_n$ to denote $C^{-1}A_n \leq B_n \leq CA_n$ for some constant $C > 1$. We will use c or C to represent a generic positive constant which may vary from line to line.

In the methodology and theory development, we assume that we have access to i.i.d. observations from class 0, $\mathcal{S}^0 = \{X^0_1, \cdots, X^0_{n_0}\}$, and i.i.d. observations from class 1, $\mathcal{S}^1 = \{X^1_1, \cdots, X^1_{n_1}\}$, where the sample sizes n_0 and n_1 are non-random positive integers. Moreover, the observations in $\mathcal{S}^0$ and $\mathcal{S}^1$ are independent. We also assume the following assumption unless specified otherwise.

Assumption 1 *(i) (On feature dimensionality and sample sizes): the dimension of features p and the sample sizes of the two classes n_0, n_1 satisfy $n_0/n > c_0, n_1/n > c_1$ for some positive* *constants c_0 and c_1, and*

$$r \equiv r_n := p/n \to r_0 \in [0,1)$$

as the sample size $n = n_0 + n_1 \to \infty$.
(ii) (On Mahalanobis distance): we assume that

$$\varDelta_d := \boldsymbol{\mu}_d^\top \Sigma^{-1} \boldsymbol{\mu}_d \geq c_2 \tag{2.2}$$

for some positive constant $c_2 > 0$.

Assumption 1 is quite natural and almost minimal to the LDA model about Σ, $\boldsymbol{\mu}^0$, and $\boldsymbol{\mu}^1$. First, our theory strongly depends on the analysis of population and sample covariance matrices. To make the inverse sample covariance matrix $\widehat{\Sigma}^{-1}$ well-defined, we have to restrict the ratio p/n strictly smaller than 1. Moreover, the sample size for either class needs to be comparable to the total sample size; otherwise, the class with a negligible sample size would be treated as noises. Second, since the Mahalanobis distance characterizes the difference between the two classes, we adopt the common regularity condition in the literature that it is bounded from below by some positive constant.

To create a sample-based classifier, the most straightforward strategy is to replace the unknown parameters in (2.1) with their sample counterparts. However, this strategy is not appropriate for our inquiry for two reasons: (i) it is well-known that direct substitutions can result in inaccurate estimates when $p/n \to r_0 \in (0,1)$; (ii) we aim for a high probability control on the type I error of the constructed classifier, and for that goal, a naive plug-in will not even work for fixed feature dimensionality. These two concerns demand that delicate refinements and corrections be made to the sample counterparts.

Before diving into the classifier construction in the next section, we introduce the notations for sample covariance matrix $\widehat{\Sigma}$ and sample mean vectors $\hat{\boldsymbol{\mu}}^a$, $a = 0,1$, and express them in forms that are more amenable in our analysis. Recall that

$$\widehat{\Sigma} = \frac{1}{n_0 + n_1 - 2} \sum_{a=0,1} \sum_{i=1}^{n_a} (X_i^a - \hat{\boldsymbol{\mu}}^a)(X_i^a - \hat{\boldsymbol{\mu}}^a)^\top, \quad \hat{\boldsymbol{\mu}}^a = \frac{1}{n_a}\left(X_1^a + \ldots + X_{n_a}^a\right), \quad a = 0,1.$$

We set the p by n data matrix by $X = (x_{ij})_{p,n} := (X^0, X^1)$, where

$$X^a := \frac{1}{(np)^{1/4}} \Sigma^{-\frac{1}{2}} (X_1^a - \boldsymbol{\mu}^a, \cdots, X_{n_a}^a - \boldsymbol{\mu}^a), \quad a = 0,1.$$

Note that all entries in the $p \times n$ matrix X are i.i.d. Gaussian with mean 0 and variance $1/\sqrt{np}$. The scaling $1/(np)^{1/4}$ is to ensure that the spectrum of $XX^\top$ has asymptotically a fixed diameter, making it a convenient choice for technical derivations. We define two unit column vectors of dimension n:

$$\mathbf{e}_0 := \frac{1}{\sqrt{n_0}} (\mathbf{1}_{n_0}^\top, 0, \cdots, 0)^\top, \quad \mathbf{e}_1 := \frac{1}{\sqrt{n_1}} (0, \cdots, 0, \mathbf{1}_{n_1}^\top)^\top. \tag{2.3}$$

With the above notations, we can rewrite the sample covariance matrix $\widehat{\Sigma}$ as

$$\widehat{\Sigma} = \frac{\sqrt{np}}{n-2} \Sigma^{\frac{1}{2}} X \Big(I_n - EE^\top\Big) X^\top \Sigma^{\frac{1}{2}}, \text{ where } E := (\mathbf{e}_0, \mathbf{e}_1). \tag{2.4}$$

For the sample means, we can rewrite them as

$$\hat{\boldsymbol{\mu}}^a = \sqrt{\frac{n}{n_a}}\, r^{\frac{1}{4}}\Sigma^{\frac{1}{2}} X\mathbf{e}_a + \boldsymbol{\mu}^a\,, \qquad a = 0, 1\,. \tag{2.5}$$

Furthermore, we write the sample mean difference vector as

$$\hat{\boldsymbol{\mu}}_d := \hat{\boldsymbol{\mu}}^1 - \hat{\boldsymbol{\mu}}^0 = r^{\frac{1}{4}}\Sigma^{\frac{1}{2}} X\mathbf{v}_1 + \boldsymbol{\mu}_d\,, \qquad \text{where } \mathbf{v_1} := \begin{pmatrix} -\frac{\sqrt{n}}{n_0}\mathbf{1}_{n_0} \\ \frac{\sqrt{n}}{n_1}\mathbf{1}_{n_1} \end{pmatrix} = -\sqrt{\frac{n}{n_0}}\mathbf{e}_0 + \sqrt{\frac{n}{n_1}}\mathbf{e}_1\,. \tag{2.6}$$

3. New Classifiers and Main Theoretical Results

In this section, we propose our new NP classifier eLDA and establish its theoretical properties regarding type I and type II errors. We also construct a variant classifier feLDA for fixed feature dimensions.

To motivate the construction of eLDA, we introduce an *intermediate* level-α NP oracle

$$\tilde{\phi}^*_\alpha(x) = \mathbb{I}\Big(\widehat{A}^\top x > \sqrt{\widehat{A}^\top \Sigma \widehat{A}}\,\Phi^{-1}(1-\alpha) + \widehat{A}^\top \boldsymbol{\mu}^0\Big)\,, \tag{3.1}$$

where $\widehat{A} = \widehat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_d$ is a shorthand notation we will frequently use in this manuscript. One can easily deduce that the type I error of $\tilde{\phi}^*_\alpha(\cdot)$ in (3.1) is exactly α. Note that $\tilde{\phi}^*_\alpha(\cdot)$ involves unknown parameters Σ and $\boldsymbol{\mu}^0$, so it is not a sample-based classifier. However, it is still of interest to compare the type II error of $\tilde{\phi}^*_\alpha(\cdot)$ to that of the level-α NP oracle in (2.1).

Lemma 1 *Let $\tilde{\phi}^*_\alpha(\cdot)$ be defined in (3.1). Under Assumption 1, the type I error of $\tilde{\phi}^*_\alpha(\cdot)$ is exactly α, i.e., $R_0(\tilde{\phi}^*_\alpha) = \alpha$. Further if $r = p/n \to 0$, then for any $\varepsilon \in (0, 1/2)$, when $n > n(\varepsilon)$, we have with probability at least $1 - n^{-1}$, the type II error satisfies*

$$R_1(\tilde{\phi}^*_\alpha) - R_1(\phi^*_\alpha) \le C\Big(r + n^{-\frac{1}{2}+\varepsilon}\Big)\sqrt{\varDelta_d}\,\exp\Big(-\frac{c\varDelta_d}{2}\Big)$$

for some constants $C, c > 0$, where C may depend on $c_{0,1,2}$ and α, and $\varDelta_d$ is defined in (2.2).

Lemma 1 indicates that $R_1(\tilde{\phi}^*_\alpha) - R_1(\phi^*_\alpha)$ goes to 0 under Assumption 1 and $p/n \to 0$. This prompts us to construct a fully sample-based classifier by modifying the unknown parts of $\tilde{\phi}^*_\alpha(\cdot)$. Towards that, we denote the threshold of $\widehat{A}^\top x$ in $\tilde{\phi}^*_\alpha(\cdot)$ by

$$F(\Sigma, \boldsymbol{\mu}^0) := \sqrt{\widehat{A}^\top \Sigma \widehat{A}}\,\Phi^{-1}(1-\alpha) + \widehat{A}^\top \boldsymbol{\mu}^0\,, \tag{3.2}$$

and denote a sample-based estimate of $F(\Sigma, \boldsymbol{\mu}^0)$ by $\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0)$, whose exact form will be introduced shortly. By studying the difference between $F(\Sigma, \boldsymbol{\mu}^0)$ and $\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0)$, we will construct a statistic $\widehat{C}^p_\alpha$ based on $\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0)$ (where the superscript p stands for parametric) that is slightly larger than $F(\Sigma, \boldsymbol{\mu}^0)$ with high probability. The proposed classifier eLDA will then be defined by replacing $F(\Sigma, \boldsymbol{\mu}^0)$ in (3.1) with $\widehat{C}^p_\alpha$.

Concretely, suppose we hope that the probability of type I error of `eLDA` no larger than α is at least around $1-\delta$, for some small given constant $\delta \in (0,1)$. We define

$$\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0) := \frac{\sqrt{\widehat{A}^\top \widehat{\Sigma} \widehat{A}}}{1-r}\, \Phi^{-1}(1-\alpha) + \widehat{A}^\top \hat{\boldsymbol{\mu}}^0 - \sqrt{\frac{n}{n_0}} \frac{r}{1-r} \mathbf{v}_1^\top \mathbf{e}_0 \,, \tag{3.3}$$

$$\widehat{C}_\alpha^p := \widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0) + \sqrt{\frac{\big((1-r)\widehat{A}^\top \widehat{\Sigma} \widehat{A} - r\|\mathbf{v}_1\|^2\big)\widehat{V}}{n}}\, \Phi^{-1}(1-\delta) \,, \tag{3.4}$$

in which $\widehat{V} = \sum_{i=1}^3 \widehat{V}_i$ and

$$\begin{aligned}
&\widehat{V}_1 := \big((1-r)\widehat{A}^\top \widehat{\Sigma} \widehat{A} - r\|\mathbf{v}_1\|^2\big) \mathsf{C}^2 \Phi_\alpha^2 \frac{2(1+r)}{(1-r)^7} \,, \\
&\widehat{V}_2 := \mathsf{C}^2 \Phi_\alpha^2 \|\mathbf{v}_1\|^2 \frac{4r(1+r)}{(1-r)^7} + \frac{n}{n_0(1-r)^3} + 2\mathsf{C}\Phi_\alpha \|\mathbf{v}_1\| \sqrt{\frac{n_1}{n_0}} \frac{2r}{(1-r)^5} \,, \\
&\widehat{V}_3 := \frac{\|\mathbf{v}_1\|^2}{(1-r)\widehat{A}^\top \widehat{\Sigma} \widehat{A} - r\|\mathbf{v}_1\|^2} \Big(\mathsf{C}^2 \Phi_\alpha^2 \|\mathbf{v}_1\|^2 \frac{2r^2(1+r)}{(1-r)^7} + \frac{(n+n_1)r}{n_0(1-r)^3} + 2\mathsf{C}\Phi_\alpha \|\mathbf{v}_1\| \sqrt{\frac{n_1}{n_0}} \frac{2r^2}{(1-r)^5} \Big) \,,
\end{aligned} \tag{3.5}$$

where $\mathsf{C} := (1-r)(\hat{\boldsymbol{\mu}}_d^\top \widehat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}_d)^{-\frac{1}{2}}/2$, $\Phi_\alpha := \Phi^{-1}(1-\alpha)$ and $\widehat{A} := \widehat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_d$.

To construct $\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0)$ and $\widehat{C}_\alpha^p$, we start with the analysis of the quadratic forms $\widehat{A}^\top \Sigma \widehat{A}$, $\widehat{A}^\top \boldsymbol{\mu}^0$ as well as their fully plug-in counterparts $\widehat{A}^\top \widehat{\Sigma} \widehat{A}$, $\widehat{A}^\top \hat{\boldsymbol{\mu}}^0$. Once we obtain their expansions (Lemma 3) and compare their leading terms, we have the estimator $\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0)$ in (3.3). However, only having $\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0)$ close to $F(\Sigma, \boldsymbol{\mu}^0)$ in (3.2) is not enough for the construction of an NP classifier. Note that the sign of $\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0) - F(\Sigma, \boldsymbol{\mu}^0)$ is uncertain. If the error is negative, directly using $\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0)$ as the threshold can actually push the type I error above α, which violates our top priority to maintain the type I error below the pre-specified level α. To address this issue, we further study the asymptotic distribution of $\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0) - F(\Sigma, \boldsymbol{\mu}^0)$ and involve a proper quantile of this asymptotic distribution in the threshold. This gives the expression of $\widehat{C}_\alpha^p$ in (3.4). By this construction, we see that $\widehat{C}_\alpha^p$ is larger than $F(\Sigma, \boldsymbol{\mu}^0)$ with high probability so that the type I error will be maintained below α with high probability. Thanks to the closeness of $\widehat{C}_\alpha^p$ to $F(\Sigma, \boldsymbol{\mu}^0)$, the excess type II error of our new classifier `eLDA` shall be close to that of $\tilde{\phi}_\alpha^*(\cdot)$. Further by Lemma 1, we shall expect the excess type II error of `eLDA` be close to that of $\phi_\alpha^*(\cdot)$, at least when $p/n \to 0$
.

Now with the above definitions, we formally introduce the new NP classifier `eLDA`:

$$\hat{\phi}_\alpha(x) = \mathbb{I}\left(\widehat{A}^\top x > \widehat{C}_\alpha^p\right) \,,$$

whose theoretical properties are described in the next theorem.

Theorem 1 *Suppose that Assumption 1 holds. For any $\alpha, \delta \in (0,1)$, let $\hat{\phi}_\alpha(x) = \mathbb{I}\Big(\widehat{A}^\top x > \widehat{C}_\alpha^p\Big)$, where $\widehat{C}_\alpha^p$ is defined in (3.4). Recall Δ_d in (2.2). Then there exist a positive constant C_1, such that for any $\varepsilon \in (0, 1/2)$, when $n > n(\varepsilon)$, it holds with probability at least $1 - \delta - C_1 n^{-\frac{1}{2}+\varepsilon}$,*

(i) the type I error satisfies: $R_0(\hat{\phi}_\alpha) \leq \alpha$*;*
(ii) for the type II error, if $r = p/n \to 0$*,*

$$R_1(\hat{\phi}_\alpha) - R_1(\phi^*_\alpha) \leq C\Big(r + n^{-\frac{1}{2}+\varepsilon}\Big)\sqrt{\Delta_d}\,\exp\Big(-\frac{c\Delta_d}{2}\Big)\,, \tag{3.6}$$

for some constants $C, c > 0$*, where* C *may depend on* $c_{0,1,2}$ *and* α*; if* $r = p/n \to r_0 \in (0,1)$*,*

$$\mathcal{L} \leq R_1(\hat{\phi}_\alpha) - R_1(\phi^*_\alpha) \leq \mathcal{U}\,,$$

where

$$\mathcal{L} := \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{1}{2}\big(\Phi_\alpha - \delta_1\sqrt{\Delta_d}\,\big)^2\Big)(1 - \sqrt{1-r} - n^{-\frac{1}{2}+\varepsilon})\sqrt{\Delta_d}\,,$$
$$\mathcal{U} := \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{1}{2}\big(\Phi_\alpha - \delta_2\sqrt{\Delta_d}\,\big)^2\Big)\Big(1 - \frac{\sqrt{1-r}}{\sigma} + n^{-\frac{1}{2}+\varepsilon}\Big)\sqrt{\Delta_d}\,,$$

for $\Phi_\alpha = \Phi^{-1}(1-\alpha)$*, and some* $\sigma > 1$*,* $\delta_1 \in (\sqrt{1-r}\,, 1)$*,* $\delta_2 \in (\sqrt{1-r}/\sigma\,, 1)$*.*

Remark 1 *We comment on the excess type II error in Theorem 1. When* $p/n \to 0$*, the upper bound can be further bounded from above by a simpler form* $C\left(r + n^{-\frac{1}{2}+\varepsilon}\right)\Delta_d^{-\beta/2}$ *for arbitrary* $\beta \geq 1$*. This simpler bound clearly implies that if* $\Delta_d = O(1)$*, the excess type II error goes to 0, while if* Δ_d *diverges, the excess type II error would tend to* 0 *at a faster rate compared to the bounded* Δ_d *situation. In contrast, when* $p/n \to r_0 \in (0,1)$*, we provide explicit forms for both upper and lower bounds of the excess type II error. One can read from the lower bound* $\mathcal{L}$ *that if* Δ_d *is of constant order, the excess type II error will not decay to* 0 *since* $\mathcal{L} \asymp 1$*. Nevertheless, if* Δ_d *diverges, then* $\mathcal{U} \to 0$ *and* `eLDA` *achieves diminishing excess type II error. In addition, our Assumption 1 coincides with the previous margin assumption and detection condition (Tong, 2013; Zhao et al., 2016; Tong et al., 2020) for an NP classifier to achieve a diminishing excess type II error. The detailed discussion can be found in Appendix A.*

Next we develop `feLDA`, a variant of `eLDA`, for bounded (or fixed) feature dimensionality p. In this case, thanks to $r = O(1/n)$, we can actually simplify `eLDA`. Concretely, let $\widetilde{V} = \Phi_\alpha^2/2 + n/n_0$. Further define

$$\widetilde{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0) := \sqrt{\widehat{A}^\top\widehat{\Sigma}\widehat{A}}\,\Phi^{-1}(1-\alpha) + \widehat{A}^\top\hat{\boldsymbol{\mu}}^0\,, \tag{3.7}$$

$$\widetilde{C}^p_\alpha := \widetilde{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0) + \sqrt{\widehat{A}^\top\widehat{\Sigma}\widehat{A}}\sqrt{\frac{\widetilde{V}}{n}}\,\Phi^{-1}(1-\delta)\,. \tag{3.8}$$

Then, we can define an NP classifier `feLDA`: $\hat{\phi}^f_\alpha(x) = \mathbb{1}\Big(\widehat{A}^\top x > \widetilde{C}^p_\alpha\Big)$, and we have the following corollary.

Corollary 1 *Suppose that Assumption 1 holds. Further, we assume that* $p = O(1)$*. For* $\alpha, \delta \in (0,1)$*, let* $\hat{\phi}^f_\alpha(x) = \mathbb{1}\Big(\widehat{A}^\top x > \widetilde{C}^p_\alpha\Big)$*, where* $\widetilde{C}^p_\alpha$ *is defined in* (3.8)*. Then there exist a* *constant C_1, such that for any $\varepsilon \in (0, 1/2)$, when $n > n(\varepsilon)$, it holds with probability at least $1 - \delta - C_1 n^{-\frac{1}{2}+\varepsilon}$,*

$$R_0(\hat{\phi}^f_\alpha) \le \alpha\,, \qquad \textit{and} \qquad R_1(\hat{\phi}^f_\alpha) - R_1(\phi^*_\alpha) \le C n^{-\frac{1}{2}+\varepsilon}\sqrt{\Delta_d}\,\exp\Big(-\frac{c\Delta_d}{2}\Big)$$

for some constants $C, c > 0$, where C may depend on $c_{0,1,2}$ and α, and Δ_d is defined in (2.2).

Note that there is no essential difference between `eLDA` and `feLDA`. The definitions of $\widetilde{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0)$ and $\widetilde{C}^p_\alpha$ are merely simplified counterparts of (3.3) and (3.4) by neglecting terms related to r; they are negligible due to the approximate $O(1/n)$ size of r. The proof of Corollary 1 is relegated to Appendix B.

4. Theoretical Results

4.1 Technical Preliminaries

In this section, we collect a few basic notions in random matrix theory and introduce some preliminary results that serve as technical inputs in our classifier construction process.

Recall the $p \times n$ data matrix X whose entries are i.i.d. Gaussian with mean 0, variance $1/\sqrt{np}$. We introduce its sample covariance matrix $H := XX^\top$ and the matrix $\mathcal{H} := X^\top X$ which has the same non-trivial eigenvalues as H. Their Green functions are defined by

$$\mathcal{G}_1(z) := (H - z)^{-1}\,, \quad \mathcal{G}_2(z) := (\mathcal{H} - z)^{-1}\,, \qquad z \in \mathbb{C}^+ := \{x + iy \mid y > 0; x, y \in \mathbb{R}\}\,.$$

Besides, we denote the normalized traces of $\mathcal{G}_1(z)$ and $\mathcal{G}_2(z)$ by

$$m_{1n}(z) := \frac{1}{p}\mathrm{Tr}\mathcal{G}_1(z) = \int \frac{1}{x - z}\, dF_{1n}(x)\,, \quad m_{2n}(z) := \frac{1}{n}\mathrm{Tr}\mathcal{G}_2(z) = \int \frac{1}{x - z}\, dF_{2n}(x)\,,$$

where $F_{1n}(x)$, $F_{2n}(x)$ are the empirical spectral distributions of H and $\mathcal{H}$ respectively, i.e.,

$$F_{1n}(x) := \frac{1}{p}\sum_{i=1}^{p} \mathbb{I}(\lambda_i(H) \le x)\,, \quad F_{2n}(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(\lambda_i(\mathcal{H}) \le x)\,.$$

Here we used $\lambda_i(H)$ and $\lambda_i(\mathcal{H})$ to denote the i-th largest eigenvalue of H and $\mathcal{H}$, respectively. Observe that $\lambda_i(H) = \lambda_i(\mathcal{H})$ for $i = 1, \cdots, p$.

It is well-known that $F_{1n}(x)$ and $F_{2n}(x)$ converge weakly (a.s.) to the *Marchenko-Pastur* laws $\nu_{\mathrm{MP},1}$ and $\nu_{\mathrm{MP},2}$ (respectively) given below

$$\begin{aligned} \nu_{\mathrm{MP},1}(\mathrm{d}x) &:= \frac{1}{2\pi x\sqrt{r}}\sqrt{\big((\lambda_+ - x)(x - \lambda_-)\big)_+}\mathrm{d}x + (1 - \frac{1}{r})_+\delta(\mathrm{d}x)\,, \\ \nu_{\mathrm{MP},2}(\mathrm{d}x) &:= \frac{\sqrt{r}}{2\pi x}\sqrt{\big((\lambda_+ - x)(x - \lambda_-)\big)_+}\mathrm{d}x + (1 - r)_+\delta(\mathrm{d}x)\,, \end{aligned} \tag{4.1}$$

where $\lambda_\pm := \sqrt{r} + 1/\sqrt{r} \pm 2$. Note that here the parameter r may be n-dependent. Hence, the weak convergence (a.s.) shall be understood as $\int g(x)\mathrm{d}F_{an}(x) - \int g(x)\nu_{\mathrm{MP},a}(\mathrm{d}x) \xrightarrow{a.s.} 0$ for any given bounded continuous function $g : \mathbb{R} \to \mathbb{R}$, for $a = 1, 2$. Note that m_{1n} and m_{2n}

can be regarded as the Stieltjes transforms of F_{1n} and F_{2n}, respectively. We further define their deterministic counterparts, i.e., Stieltjes transforms of $\nu_{\mathrm{MP},1}, \nu_{\mathrm{MP},2}$, by $m_1(z), m_2(z)$, respectively, i.e., $m_a(z) := \int (x-z)^{-1} \nu_{\mathrm{MP},a}(\mathrm{d}x)$, for $a = 1, 2$. From the definition (4.1), it is straightforward to derive

$$\begin{aligned} m_1(z) &= \frac{r^{-1/2} - r^{1/2} - z + \mathrm{i}\sqrt{(\lambda_+ - z)(z - \lambda_-)}}{2r^{1/2}z}\,, \\ m_2(z) &= \frac{r^{1/2} - r^{-1/2} - z + \mathrm{i}\sqrt{(\lambda_+ - z)(z - \lambda_-)}}{2r^{-1/2}z}\,, \end{aligned} \tag{4.2}$$

where the square root is taken with a branch cut on the negative real axis. Equivalently, we can also characterize $m_1(z), m_2(z)$ as the unique solutions from $\mathbb{C}^+$ to $\mathbb{C}^+$ to the equations

$$zr^{1/2}m_1^2 + [z - r^{-1/2} + r^{1/2}]m_1 + 1 = 0\,, \quad zr^{-1/2}m_2^2 + [z - r^{1/2} + r^{-1/2}]m_2 + 1 = 0\,. \tag{4.3}$$

In later discussions, we need the estimates of the quadratic forms of Green functions. Towards that, we define the notion *stochastic domination* which was initially introduced in (Erdős et al., 2013). It provides a precise statement of the form "X_N is bounded by Y_N up to a small power of N with high probability".

Definition 1 *(Stochastic domination) Let*

$$\mathsf{X} = \big(\mathsf{X}_N(u) : N \in \mathbb{N}, u \in U_N\big) \quad \text{and} \quad \mathsf{Y} = \big(\mathsf{Y}_N(u) : N \in \mathbb{N}, u \in U_N\big)$$

be two families of random variables, Y *is nonnegative, and* U_N *is a possibly* N*-dependent parameter set. We say that* X *is stochastically dominated by* Y*, uniformly in* u*, if for all small* $\varrho > 0$ *and large* $\phi > 0$*, we have*

$$\sup_{u \in U_N} \mathbb{P}\big(|\mathsf{X}_N(u)| > N^{\varrho}\mathsf{Y}_N(u)\big) \le N^{-\phi}$$

for large $N \ge N_0(\varrho, \phi)$*. Throughout the paper, we use the notation* $\mathsf{X} = O_\prec(\mathsf{Y})$ *or* $\mathsf{X} \prec \mathsf{Y}$ *when* X *is stochastically dominated by* Y *uniformly in* u*. Note that in the special case when* X *and* Y *are deterministic,* $\mathsf{X} \prec \mathsf{Y}$ *means for any given* $\varrho > 0$*,* $|\mathsf{X}_N(u)| \le N^{\varrho}\mathsf{Y}_N(u)$ *uniformly in* u*, for all sufficiently large* $N \ge N_0(\varrho)$*.*

Definition 2 *Two sequences of random vectors,* $\mathsf{X}_N \in \mathbb{R}^k$ *and* $\mathsf{Y}_N \in \mathbb{R}^k$*,* $N \ge 1$*, are* asymptotically equal in distribution*, denoted as* $\mathsf{X}_N \simeq \mathsf{Y}_N$*, if they are tight and satisfy* $\lim_{N\to\infty} \big(\mathbb{E}f(\mathsf{X}_N) - \mathbb{E}f(\mathsf{Y}_N)\big) = 0$*, for any bounded continuous function* $f : \mathbb{R}^k \to \mathbb{R}$*.*

Further, we introduce a basic lemma based on Definition 1.

Lemma 2 *Let* $\mathsf{X}_i = (\mathsf{X}_{N,i}(u) : N \in \mathbb{N},\ u \in U_N)$ *,* $\mathsf{Y}_i = (\mathsf{Y}_{N,i}(u) : N \in \mathbb{N},\ u \in U_N)$*,* $i = 1, 2$*, be families of random variables, where* $\mathsf{Y}_i, i = 1, 2$*, are nonnegative, and* U_N *is a possibly* N*-dependent parameter set. Let* $\Phi = (\Phi_N(u) : N \in \mathbb{N},\ u \in U_N)$ *be a family of deterministic nonnegative quantities. We have the following results:*

(i) If $\mathsf{X}_1 \prec \mathsf{Y}_1$ *and* $\mathsf{X}_2 \prec \mathsf{Y}_2$ *then* $\mathsf{X}_1 + \mathsf{X}_2 \prec \mathsf{Y}_1 + \mathsf{Y}_2$ *and* $\mathsf{X}_1\mathsf{X}_2 \prec \mathsf{Y}_1\mathsf{Y}_2$*.*

(ii) Suppose $\mathsf{X}_1 \prec \Phi$*, and there exists a constant* $C > 0$ *such that* $|\mathsf{X}_{N,1}(u)| \le N^C$ *a.s. and* $\Phi_N(u) \ge N^{-C}$ *uniformly in* u *for all sufficiently large* N*. Then* $\mathbb{E}\mathsf{X}_1 \prec \Phi$*.*

We introduce the following domain. For a small fixed τ, we define

$$\mathcal{D}^0 \equiv \mathcal{D}(\tau)^0 := \{z \in \mathbb{C}^+ : -\tau < \Re z < \tau, 0 < \Im z \le \tau^{-1}\} . \tag{4.4}$$

Conventionally, for $a = 1, 2$, we use $\mathcal{G}_a^\ell$ and $\mathcal{G}_a^{(\ell)}$ to represent ℓ-th power of $\mathcal{G}_a$ and the ℓ-th derivative of $\mathcal{G}_a$ with respect to z, respectively. With these notations, we introduce the following proposition which is known as local laws and shall be regarded as slight adaptations of the results in (Bloemendal et al., 2014), in the Gaussian case.

Proposition 1 *Let $\tau > 0$ be a small but fixed constant. Under Assumption 1, for any given $l \in \mathbb{N}$, we have*

$$\left|\big(\mathcal{G}_1^{(l)}(z)\big)_{ij} - m_1^{(l)}(z)\delta_{ij}\right| \prec n^{-\frac{1}{2}} r^{\frac{1+l}{2}} , \quad \left|\big(z\mathcal{G}_2(z)\big)^{(l)}_{i'j'} - \big(zm_2(z)\big)^{(l)}\delta_{i'j'}\right| \prec n^{-\frac{1}{2}} r^{\frac{1+l}{2}} , \tag{4.5}$$

$$\left|\big(X^\top \mathcal{G}_1^{(l)}(z)\big)_{i'i}\right| \prec n^{-\frac{1}{2}} r^{\frac{1}{4}+\frac{l}{2}} , \quad \left|\Big(X\big(z\mathcal{G}_2(z)\big)^{(l)}\Big)_{ii'}\right| \prec n^{-\frac{1}{2}} r^{-\frac{1}{4}+\frac{l}{2}} , \tag{4.6}$$

$$\left|m_{1n}^{(l)}(z) - m_1^{(l)}(z)\right| \prec n^{-1} r^{\frac{l}{2}} , \quad \left|\big(zm_{2n}(z)\big)^{(l)} - \big(zm_2(z)\big)^{(l)}\right| \prec n^{-1} r^{\frac{1+l}{2}} , \tag{4.7}$$

uniformly in $z \in \mathcal{D}^0$ and for any $i, j \in \{1, \cdots, p\}$ and $i', j' \in \{1, \cdots, n\}$. For $l = 0$, the second estimates in (4.6) and (4.7) can be improved to

$$\left|\Big(X\big(z\mathcal{G}_2(z)\big)\Big)_{ii'}\right| \prec n^{-\frac{1}{2}} r^{\frac{1}{4}} , \quad \left|\big(zm_{2n}(z)\big) - \big(zm_2(z)\big)\right| \prec n^{-1} r . \tag{4.8}$$

Remark 2 *By the orthogonal invariance of Gaussian random matrix, we get from Proposition 1 that for $\mathbf{u}, \mathbf{v}$, any complex deterministic unit vectors of proper dimensions,*

$$|\langle \mathbf{u}, \mathcal{G}_1^{(l)}(z)\mathbf{v}\rangle - m_1^{(l)}(z)\langle \mathbf{u}, \mathbf{v}\rangle| \prec n^{-\frac{1}{2}} r^{\frac{1+l}{2}}, \quad |\langle \mathbf{u}, \big(z\mathcal{G}_2(z)\big)^{(l)}\mathbf{v}\rangle - \big(zm_2(z)\big)^{(l)}\langle \mathbf{u}, \mathbf{v}\rangle| \prec n^{-\frac{1}{2}} r^{\frac{1+l}{2}}, \tag{4.9}$$

$$|\langle \mathbf{u}, X^\top \mathcal{G}_1^{(l)}(z)\mathbf{v}\rangle| \prec n^{-\frac{1}{2}} r^{\frac{1}{4}+\frac{l}{2}}, \quad |\langle \mathbf{u}, X\big(z\mathcal{G}_2(z)\big)^{(l)}\mathbf{v}\rangle| \prec n^{-\frac{1}{2}} r^{\frac{1}{4}+\frac{l}{2}}, \tag{4.10}$$

uniformly for $z \in \mathcal{D}^0$. We further remark that the estimates above and the ones in Proposition 1 also hold at $z = 0$ with error bounds unchanged by the Lipschitz continuity of $\mathcal{G}_1, z\mathcal{G}_2(z)$, $m_1(z)$, and $zm_2(z)$. And we will use (4.7), (4.9), and (4.10) frequently in technical proofs not only for $z \in \mathcal{D}^0$ but also at $z = 0$.

4.2 Key Technical Results

In this section, we prove our main theorem, i.e., Theorem 1. To streamline the proof, we first present two technical results and their proof sketches.

Lemma 3 *Suppose that Assumption 1 holds. Recall the definition of Δ_d in (2.2). Let $\widehat{A} = \widehat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_d$, then we have*

$$\widehat{A}^\top \Sigma \widehat{A} = \frac{r}{(1-r)^3}\|\mathbf{v}_1\|^2 + \frac{1}{(1-r)^3}\,\Delta_d + O_\prec\big(n^{-\frac{1}{2}}\Delta_d\big) , \tag{4.11}$$

$$\widehat{A}^\top \widehat{\Sigma} \widehat{A} = \frac{r}{1-r}\|\mathbf{v}_1\|^2 + \frac{1}{1-r}\,\Delta_d + O_\prec\big(n^{-\frac{1}{2}}\Delta_d\big) , \tag{4.12}$$

$$\widehat{A}^\top \boldsymbol{\mu}_d = \frac{1}{1-r}\Delta_d + O_\prec\big(n^{-\frac{1}{2}}\Delta_d\big) , \tag{4.13}$$

$$\widehat{A}^\top \hat{\boldsymbol{\mu}}^0 - \widehat{A}^\top \boldsymbol{\mu}^0 = \sqrt{\frac{n}{n_0}}\frac{r}{1-r}\mathbf{v}_1^\top \mathbf{e}_0 + O_\prec\big(n_0^{-\frac{1}{2}}\Delta_d^{\frac{1}{2}}\big) . \tag{4.14}$$

Moreover, counterparts of (4.14) also hold if the triple $(\boldsymbol{\mu}^0, \hat{\boldsymbol{\mu}}^0, \sqrt{n/n_0}\,\mathbf{e}_0)$ *is replaced by* $(\boldsymbol{\mu}^1, \hat{\boldsymbol{\mu}}^1, \sqrt{n/n_1}\,\mathbf{e}_1)$ *or* $(\boldsymbol{\mu}_d, \hat{\boldsymbol{\mu}}_d, \mathbf{v}_1)$*.*

Remark 3 *Lemma 3 hints that we can use* $\widehat{A}^\top\widehat{\Sigma}\widehat{A}/(1-r)^2$ *to estimate* $\widehat{A}^\top\Sigma\widehat{A}$ *and use* $\widehat{A}^\top\hat{\boldsymbol{\mu}}^0 - \sqrt{\frac{n}{n_0}}\frac{r}{1-r}\mathbf{v}_1^\top\mathbf{e}_0$ *to approximate* $\widehat{A}^\top\boldsymbol{\mu}^0$*. Therefore, we construct* $\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0)$*, whose definition is explicitly given in (3.3). Moreover, when* p *is fixed, i.e.,* $r = O(1/n)$*, we get the following simplified estimates*

$$\widehat{A}^\top\Sigma\widehat{A} = \mathit{\Delta}_d + O_\prec(n^{-\frac{1}{2}}\mathit{\Delta}_d)\,, \quad \widehat{A}^\top\widehat{\Sigma}\widehat{A} = \mathit{\Delta}_d + O_\prec(n^{-\frac{1}{2}}\mathit{\Delta}_d)\,, \tag{4.15}$$

$$\widehat{A}^\top\boldsymbol{\mu}_d = \mathit{\Delta}_d + O_\prec(n^{-\frac{1}{2}}\mathit{\Delta}_d)\,, \quad \widehat{A}^\top\hat{\boldsymbol{\mu}}^0 - \widehat{A}^\top\boldsymbol{\mu}^0 = O_\prec(n_0^{-\frac{1}{2}}\mathit{\Delta}_d^{\frac{1}{2}})\,. \tag{4.16}$$

We provide a proof sketch of Lemma 3, while a formal proof is presented in the Supplementary Materials. Our starting point is to expand $\widehat{\Sigma}^{-1}$ in terms of Green function $\mathcal{G}_1(z) = (XX^\top - z)^{-1}$ at $z = 0$ since all the quadratic forms in Lemma 3 can be rewritten as certain quadratic forms of $\widehat{\Sigma}^{-1}$ according to the representations (2.4)-(2.6). Working with Green functions makes the analysis much easier due to the useful estimates in local laws, i.e., Proposition 1 and its variants (4.9), (4.10). In this expansion, we will need some elementary linear algebra (e.g., Woodbury matrix identity) to compute matrix inverse and local laws (4.7), (4.9) and (4.10) to estimate the error terms. Next, with the expansion of $\widehat{\Sigma}^{-1}$ plugged in, all the quadratic forms we want to study in Lemma 3 can be further simplified to linear combinations of quadratic forms of $\mathcal{G}_1^a(0)$, $\mathcal{G}_1^a(0)X$, and $X^\top\mathcal{G}_1^a(0)X$, for $a = 1, 2$. Then, further derivations with the aid of local laws (4.7), (4.9) and (4.10) lead to the ultimate expressions. All these derivations only need the first order expansion since we focus on the leading terms.

Next, we describe the difference between $\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0)$ and $F(\Sigma, \boldsymbol{\mu}^0)$ by a quantitative CLT.

Proposition 2 *Let* $F(\Sigma, \boldsymbol{\mu}^0)$ *and* $\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0)$ *be defined in (3.2) and (3.3), respectively. Under Assumption 1, we have*

$$\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0) - F(\Sigma, \boldsymbol{\mu}^0) = \frac{\sqrt{(1-r)\hat{\boldsymbol{\mu}}_d^\top\widehat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_d - \frac{n^2 r}{n_0 n_1}}}{\sqrt{n}}\,\mathit{\Theta}_\alpha + O_\prec\Big(n^{-1}\big(r^{\frac{1}{2}} + \mathit{\Delta}_d^{\frac{1}{2}}\big)\Big)\,, \tag{4.17}$$

and the random part $\mathit{\Theta}_\alpha$ *satisfies*

$$\mathit{\Theta}_\alpha \simeq \mathcal{N}(0, \widehat{V})\,,$$

where $\widehat{V}$ *was defined in (3.5). Furthermore, the convergence rate of* $\mathit{\Theta}_\alpha$ *to* $\mathcal{N}(0, \widehat{V})$ *is* $O_\prec(n^{-1/2})$ *in Kolmogorov-Smirnov distance, i.e.,* $\sup_t\Big|\mathbb{P}\big(\Theta_\alpha \le t\big) - \mathbb{P}\big(\mathcal{N}(0, \widehat{V}) \le t\big)\Big| \prec n^{-1/2}$ *, where we simply use* $\mathcal{N}(0, \widehat{V})$ *to denote a random variable with distribution* $\mathcal{N}(0, \widehat{V})$*.*

We state the sketch of the proof of Proposition 2 as follows. First, we express $\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0) - F(\Sigma, \boldsymbol{\mu}^0)$ in terms of Green functions $\mathcal{G}_1(z) = (XX^\top - z)^{-1}$ and $(z\mathcal{G}_2(z)) = z(X^\top X - z)^{-1}$ at $z = 0$ (Lemma D.1 in Appendix D). Different from the derivations of the expansions of the quadratic forms in Lemma 3, here we need to do second order expansions for $\widehat{\Sigma}^{-1}$ and quadratic forms of $\mathcal{G}_1^a(0)$, $\mathcal{G}_1^a(0)X$ and $X^\top\mathcal{G}_1^a(0)X$, for $a = 1, 2$. Because the leading terms of $\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0)$ and $F(\Sigma, \boldsymbol{\mu}^0)$ cancel out with each other due to their definitions and Lemma 3, higher order terms are needed to study the asymptotic distribution. The error terms in the expansions can be estimated with the help of local laws (4.7), (4.9) and (4.10). It turns out that the leading terms of $\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0) - F(\Sigma, \boldsymbol{\mu}^0)$ in Lemma D.1 are given by linear combinations of certain quadratic forms of $\mathcal{G}_1^{(\ell)} - m_1^{(\ell)}$, $(z\mathcal{G}_2)^{(\ell)} - (zm_2)^{(\ell)}$ and $\mathcal{G}_1^{(\ell)} X$ where we omit the argument z in $\mathcal{G}_1$, $\mathcal{G}_2$ at $z = 0$. This inspires us to study the joint asymptotic distribution of these quadratic forms. To derive a multivariate Gaussian distribution, it is equivalent to show the asymptotically Gaussian distribution for a generic linear combination $\mathcal{P}$ of the quadratic forms appeared in the Green function representation formula; see equation (D.21) in Appendix D for the specific expression of $\mathcal{P}$. Next, we aim to derive a differential equation of the characteristic function of $\mathcal{P}$, denoted by $\phi_n(\cdot)$. Concretely, we show that for $|t| \ll n^{\frac{1}{2}}$, $\varphi_n'(t) = -Vt\varphi_n(t) + O_\prec((|t|+1)n^{-\frac{1}{2}})$, where V is some deterministic constant that indicates the variance of $\mathcal{P}$. The above estimate has two implications. First, it indicates the Gaussianity of $\mathcal{P}$. Second, applying Esseen's inequality, we can obtain its convergence rate as well. The proof of the above estimate relies on the technique of integration by parts and local laws. More details can be found in the proof of Proposition D.1 in Appendix D .

Remark 4 *In the case that p is fixed, or $r \equiv r_n = O_\prec(1/n)$, we have the simplified version of Proposition 2 where $\widetilde{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0)$ defined in* (3.7) *is involved:*

$$\widetilde{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0) - F(\Sigma, \boldsymbol{\mu}^0) = \frac{1}{\sqrt{n}} \sqrt{\hat{\boldsymbol{\mu}}_d^\top \widehat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}_d}\, \widetilde{\Theta}_\alpha + O_\prec\big(n^{-1} \Delta_d^{\frac{1}{2}}\big) , \tag{4.18}$$

and the random part $\widetilde{\Theta}_\alpha$ satisfies $\widetilde{\Theta}_\alpha \simeq \mathcal{N}(0, \widetilde{V})$ with rate $O_\prec(n^{-1/2})$. We also remark that the proof of this simplified version is similar to that of Proposition 2 by absorbing some terms containing r into the error thanks to $r = O(1/n)$. Hence, we will omit the proof.

4.3 Proof of Main Theorem

With the help of Lemma 3 and Proposition 2, we are now ready to prove the main theorem (c.f. Theorem 1).

Proof [Proof of Theorem 1] Recall that $\hat{\phi}_\alpha(x) = \mathbb{1}\Big(\widehat{A}^\top x > \widehat{C}_\alpha^p\Big)$. If we can claim that

$$\widehat{C}_\alpha^p \geq F(\Sigma, \boldsymbol{\mu}^0) \tag{4.19}$$

with high probability, then immediately, we can conclude that with high probability,

$$R_0(\hat{\phi}_\alpha) = \mathbb{P}\Big(\widehat{A}^\top \mathbf{x} > \widehat{C}_\alpha^p \Big| \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}^0, \Sigma)\Big) \leq \mathbb{P}\Big(\widehat{A}^\top \mathbf{x} > F(\Sigma, \boldsymbol{\mu}^0) \Big| \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}^0, \Sigma)\Big) = R_0(\phi_\alpha^*) = \alpha.$$

In the sequel, we establish inequality (4.19) with high probability. By the definition of $\widehat{C}_\alpha^p$ in (3.4) and the representation (4.17), we have

$$\begin{aligned} \widehat{C}_\alpha^p - F(\Sigma, \boldsymbol{\mu}^0) &= \widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0) - F(\Sigma, \boldsymbol{\mu}^0) + \sqrt{\big((1-r)\widehat{A}^\top \widehat{\Sigma} \widehat{A} - r\|\mathbf{v}_1\|^2\big)} \sqrt{\frac{\widehat{V}}{n}}\, \Phi^{-1}(1-\delta) \\ &= \frac{1}{\sqrt{n}} \sqrt{(1-r)\widehat{A}^\top \widehat{\Sigma} \widehat{A} - r\|\mathbf{v}_1\|^2}\, \Big(\Theta_\alpha - \sqrt{\widehat{V}}\, \Phi^{-1}(\delta)\Big) + O_\prec(n^{-1} \Delta_d^{\frac{1}{2}}) . \end{aligned}$$

By Proposition 2, $\mathit{\Theta}_\alpha$ is asymptotically $\mathcal{N}(0,\widehat{V})$ distributed with convergence rate $O_\prec(n^{-1/2})$. We then have for any constant $\varepsilon \in (0,\frac{1}{2})$,

$$\begin{aligned}\mathbb{P}\Big(\mathit{\Theta}_\alpha - \sqrt{\widehat{V}}\,\Phi^{-1}(\delta) > n^{-\frac{1}{2}+\varepsilon}\Big) &= \mathbb{P}\Big(\mathit{\Theta}_\alpha/\sqrt{\widehat{V}} > \Phi^{-1}(\delta) + n^{-\frac{1}{2}+\varepsilon}/\sqrt{\widehat{V}}\,\Big)\\ &\ge \mathbb{P}\Big(\mathcal{N}(0,1) > \Phi^{-1}(\delta) + n^{-\frac{1}{2}+\varepsilon}/\sqrt{\widehat{V}}\,\Big) - n^{-\frac{1}{2}+\varepsilon}\\ &= 1 - \Phi\big(\Phi^{-1}(\delta) + n^{-\frac{1}{2}+\varepsilon}/\sqrt{\widehat{V}}\,\big) - n^{-\frac{1}{2}+\varepsilon}\\ &\ge 1 - \delta - C_1 n^{-\frac{1}{2}+\varepsilon}\end{aligned}$$

for some $C_1 > 0$ and $n > n(\varepsilon)$. Here the second step is due to the convergence rate $O_\prec(n^{-1/2})$ of Θ_α; And for the last step, we used the continuity of $\Phi(\cdot)$ together with $\widehat{V} > c$ for some constant $c > 0$ following from the definition (3.5). Further we have the estimate $\sqrt{(1-r)\widehat{A}^\top\widehat{\Sigma}\widehat{A} - r\|\mathbf{v}_1\|^2} \asymp \mathit{\Delta}_d^{1/2}$ with probability at least $1 - n^{-D}$ for any $D > 0$ and $n > n(\varepsilon, D)$, which is obtained from (4.12). Thereby, we get that

$$\frac{1}{\sqrt{n}}\sqrt{(1-r)\widehat{A}^\top\widehat{\Sigma}\widehat{A} - r\|\mathbf{v}_1\|^2}\,\Big(\mathit{\Theta}_\alpha - \sqrt{\widehat{V}}\,\Phi^{-1}(\delta)\Big) \ge cn^{-1+\varepsilon}\mathit{\Delta}_d^{\frac{1}{2}}$$

for some $c > 0$, with probability at least $1 - \delta - C_1 n^{-\frac{1}{2}+\varepsilon} - n^{-D}$ when $n > n(\varepsilon, D)$. As a consequence, there exist some $C_1, C_2 > 0$ such that

$$\widehat{C}^p_\alpha - F(\Sigma, \boldsymbol{\mu}^0) > cn^{-1+\varepsilon}\big(r^{\frac{1}{2}} + \sqrt{\boldsymbol{\mu}_d^\top\Sigma^{-1}\boldsymbol{\mu}_d}\,\big) + O_\prec\Big(n^{-1}\big(r^{\frac{1}{2}} + \sqrt{\boldsymbol{\mu}_d^\top\Sigma^{-1}\boldsymbol{\mu}_d}\,\big)\big(1 + \sqrt{\frac{n}{n_0}}\,r^{\frac{1}{2}}\big)\Big) > 0$$

with probability at least $1 - \delta - C_1 n^{-\frac{1}{2}+\varepsilon} - C_2 n^{-D}$ for any $\varepsilon \in (0, 1/2)$ and $D > 0$, when $n > n(\varepsilon, D)$.

In the sequel, we proceed to prove statement (ii) regarding the type II error. Note that by definition,

$$\begin{aligned}R_1(\hat{\phi}_\alpha) &= \mathbb{P}(\hat{\phi}_\alpha(\mathbf{x}) \neq Y \big| Y = 1) = \mathbb{P}\Big(\widehat{A}^\top\mathbf{x} < \widehat{C}^p_\alpha \Big| \mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}^1, \Sigma)\Big)\\ &= \Phi\Big((\widehat{A}^\top\Sigma\widehat{A})^{-\frac{1}{2}}\big(\widehat{C}^p_\alpha - \widehat{A}^\top\boldsymbol{\mu}^1\big)\Big) = \Phi\Big(\Phi^{-1}(1-\alpha) - \frac{\widehat{A}^\top\boldsymbol{\mu}_d}{\sqrt{\widehat{A}^\top\Sigma\widehat{A}}} + O_\prec(n^{-\frac{1}{2}})\Big).\end{aligned} \tag{4.20}$$

Using the estimates in Lemma 3, if $p/n \to 0$, we further have

$$R_1(\hat{\phi}_\alpha) = \Phi\Big(\Phi^{-1}(1-\alpha) - \mathit{\Delta}_d^{\frac{1}{2}} + O_\prec\big(n^{-\frac{1}{2}}\mathit{\Delta}_d^{\frac{1}{2}}\big) + O\big(r\mathit{\Delta}_d^{\frac{1}{2}}\big)\Big).$$

Then, compared with $R_1(\phi^*_\alpha) = \Phi\Big(\Phi^{-1}(1-\alpha) - \mathit{\Delta}_d^{\frac{1}{2}}\Big)$, it is not hard to deduce that in the case of $p/n \to 0$, (3.6) holds.

In the case that $p/n \to r_0 \in (0,1)$, continuing with (4.20), we arrive at

$$R_1(\hat{\phi}_\alpha) = \Phi\Big(\Phi^{-1}(1-\alpha) - \frac{(1-r)^{-1}\mathit{\Delta}_d}{\sqrt{\frac{r}{(1-r)^3}\|\mathbf{v}_1\|^2 + \frac{1}{(1-r)^3}\mathit{\Delta}_d}} + O_\prec\big(n^{-\frac{1}{2}}\mathit{\Delta}_d^{\frac{1}{2}}\big)\Big).$$

However, in this case,

$$\frac{\sqrt{1-r}}{\sigma}\,\varDelta_d^{\frac{1}{2}} < \frac{(1-r)^{-1}\varDelta_d}{\sqrt{\frac{r}{(1-r)^3}\|\mathbf{v}_1\|^2 + \frac{1}{(1-r)^3}\varDelta_d}} < \sqrt{1-r}\,\varDelta_d^{\frac{1}{2}}\,,$$

for some $\sigma > 1$ which depends on $r\|\mathbf{v}_1\|^2/\varDelta_d$. Thereby, by some elementary computations, one shall obtain that with probability at least $1 - n^{-D}$ for $D > 0$ and $\varepsilon \in (0, 1/2)$, when $n > n(\varepsilon, D)$,

$$\begin{aligned}
R_1(\hat\phi_\alpha) - R_1(\phi^*_\alpha) &\geq \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{1}{2}\big(\Phi_\alpha - \delta_1\sqrt{\varDelta_d}\,\big)^2\Big)(1 - \sqrt{1-r} - n^{-\frac{1}{2}+\varepsilon})\sqrt{\varDelta_d}\,,\\
R_1(\hat\phi_\alpha) - R_1(\phi^*_\alpha) &\leq \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{1}{2}\big(\Phi_\alpha - \delta_2\sqrt{\varDelta_d}\big)^2\Big)\Big(1 - \frac{\sqrt{1-r}}{\sigma} + n^{-\frac{1}{2}+\varepsilon}\Big)\sqrt{\varDelta_d}\,,
\end{aligned}$$

for some $\delta_1 \in (\sqrt{1-r}, 1)$ and $\delta_2 \in (\sqrt{1-r}/\sigma, 1)$.

Combining the loss of probability for both statements together and setting $D = 1$, eventually we see that (i) and (ii) hold with probability at least $1 - \delta - C_1 n^{-\frac{1}{2}+\varepsilon}$ and hence we finished the proof of Theorem 1.

■

5. Numerical Analysis

5.1 Simulation Studies

In this section, we compare the performance of the two newly proposed classifiers `eLDA` and `feLDA` with that of five existing splitting NP methods: `pNP-LDA`, `NP-LDA`, `NP-sLDA`, `NP-svm`, and `NP-penlog`. Here `pNP-LDA` is the parametric NP classifier as discussed in Section 1, where the threshold is constructed parametrically and the base algorithm is linear discriminant analysis (LDA). The latter four methods with `NP` as the prefix use the NP umbrella algorithm to select the threshold, and the base algorithms for scoring functions are LDA, sparse linear discriminant analysis (sLDA), svm and penalized logistic regression (penlog), respectively. [1] In figures, we omit the `NP` for these four methods for concise presentation. Among the five existing methods, only `pNP-LDA` does not have sample size requirement on n_0. Thus for small n_0, we can only compare our new methods with `pNP-LDA`. For all five splitting NP classifiers, τ, the class 0 split proportion, is fixed at 0.5, and the each experiment is repeated 1,000 times.

In Example 1a, we particularly added a vanilla plug-in estimator of the oracle classifer (2.1) as a benchmark for comparison. More specifically, we replace all the population parameters with their sample counterparts:

$$\hat\phi^*_\alpha(x) = \mathbb{1}\Big((\widehat\Sigma^{-1}\widehat{\boldsymbol\mu}_d)^\top x > \sqrt{\widehat{\boldsymbol\mu}_d^\top\widehat\Sigma^{-1}\widehat{\boldsymbol\mu}_d}\,\Phi^{-1}(1-\alpha) + \widehat{\boldsymbol\mu}_d^\top\widehat\Sigma^{-1}\widehat{\boldsymbol\mu}^0\Big)\,. \tag{5.1}$$

[1] We implemented the NP umbrella algorithms using the R package `npc` with default parameters. Specifically, for `NP-svm`, `npc` adopted the "radial" kernel for analysis. It is important to note that when the data resembles a Gaussian distribution, implementing svm with a "linear" kernel may slightly enhance performance. However, even in such cases, eLDA consistently outperforms all of its competitors.

We aim to demonstrate the limitations of a simple plug-in estimator $\hat{\phi}^*_\alpha(\mathbf{x})$ in controlling the type I error under α with a desirable violation rate. This highlights the necessity of developing `eLDA` as an alternative approach.

Example 1 *The data are generated from an LDA model with common covariance matrix Σ, where Σ is set to be an AR(1) covariance matrix with $\Sigma_{ij} = 0.5^{|i-j|}$ for all i and j. $\boldsymbol{\beta}^{Bayes} = \Sigma^{-1}\boldsymbol{\mu}_d = 1.2 \times (\mathbf{1}_{p_0}, \mathbf{0}_{p-p_0})^\top$, $\boldsymbol{\mu}^0 = \mathbf{0}_p$, $p_0 = 3$. We set $\pi_0 = \pi_1 = 0.5$ and $\alpha = 0.1$. Type I and type II errors are evaluated on a test set that contains* 30,000 *observations from each class, and then we report the average over the* 1,000 *repetitions.*

(1a) $\delta = 0.1$, $p = 3$, *varying* $n_0 = n_1 \in \{20, 70, 120, 170, 220, 270, 320, 370, 500, 1000\}$

(1b) $\delta = 0.1$, $p = 3$, $n_1 = 500$, *varying* $n_0 \in \{20, 70, 120, 170, 220, 270, 320, 370, 500, 1000\}$

(1c) $\delta = 0.1$, $n_0 = n_1 = 125$, *varying* $p \in \{3, 6, 9, 12, 15, 18, 21, 24, 27, 30\}$

(1c') $\delta = 0.05$, $n_0 = n_1 = 125$, *varying* $p \in \{3, 6, 9, 12, 15, 18, 21, 24, 27, 30\}$

(1c)* $\delta = 0.01$, $n_0 = n_1 = 125$, *varying* $p \in \{3, 6, 9, 12, 15, 18, 21, 24, 27, 30\}$

(1d) $\delta = 0.1$, $n_0 = 125$, $n_1 = 500$, *varying* $p \in \{3, 6, 9, 12, 15, 18, 21, 24, 27, 30\}$

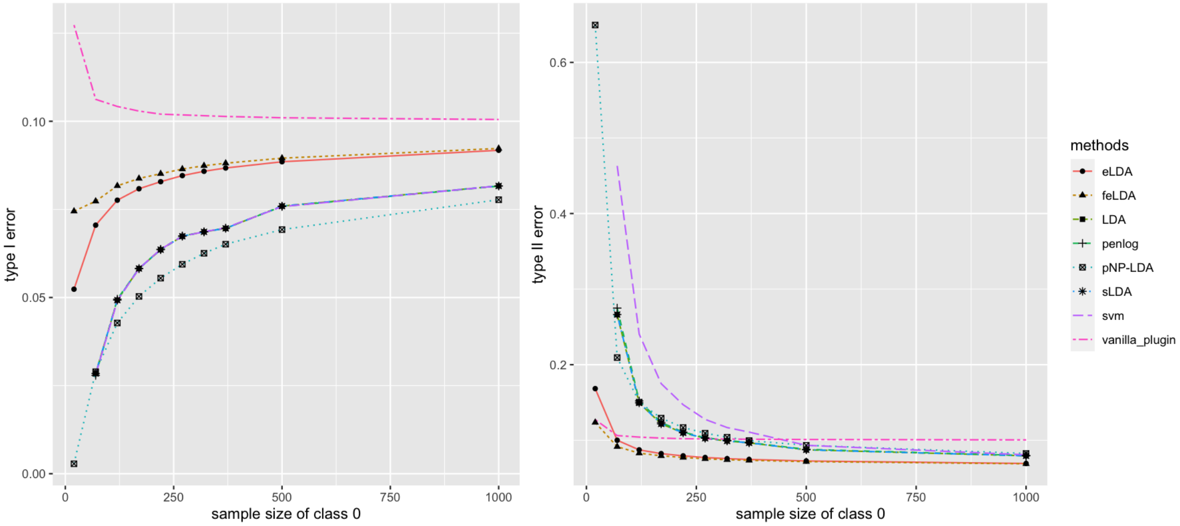
(1d') $\delta = 0.05$, $n_0 = 125$, $n_1 = 500$, *varying* $p \in \{3, 6, 9, 12, 15, 18, 21, 24, 27, 30\}$

We summarize the results for Example 1 in Figure 1, Figure 2, Table 2, Table 3 and Appendix Figures 4, 5, and 6. We discuss our findings in order.

Examples 1a and 1b share the common violation rate target $\delta = 0.1$ and low dimension $p = 3$. Their distinction comes from the two class sample sizes; Example 1a has balanced increasing sample sizes, i.e., $n_0 = n_1$, while Example 1b keeps n_1 fixed at 500, and only increases n_0. Due to the space limit, we only demonstrate the performance of Example 1a in Figure 1, in terms of type I and type II errors, and relegate the comparison between Example 1a and Example 1b to Appendix Figure 4. Notice that, for very small class 0 sample sizes $n_0 = 20$, all NP umbrella algorithm based methods (`NP-LDA`, `NP-sLDA`, `NP-svm`, and `NP-penlog`) fail to meet the minimum sample size requirement for class 0 and are not implementable, thus only the performances of `eLDA`, `feLDA` and `pNP-LDA` are available in Figure 1. Consistently across Example 1a and Example 1b, we see that 1) as n_0 increases, for all methods, the type I errors increase (but bounded above by α), and the type II errors decrease. Nevertheless, the five existing NP methods present type I errors mostly below 0.08, and are much more conservative compared to `eLDA` and `feLDA`, whose type I errors closer to 0.1; 2) in terms of type II errors, `eLDA` and `feLDA` significantly outperform the other five methods across all n_0's. Comparing Example 1b to Example 1a, keeping $n_1 = 500$ does not affect much the performance of `eLDA` and `feLDA`. However, Example 1b has aggravated the type I error performance of `pNP-LDA` for small n_0, and also the type II error performance of `NP-svm`. It is important to highlight that the `vanilla plug-in` classifier $\hat{\phi}^*_\alpha(x)$ has shown limitations in effectively controlling even the average type I error. This is particularly evident with small sample sizes (e.g., less than 250), where the type I error exceeds α by a significant margin. This observation underscores the importance of delicate threshold construction in `eLDA`.

We further summarize the observed (type I error) violation rate[2] in Table 2. The five splitting NP classifiers all have violation rates smaller than targeted $\delta = 0.1$, and share a common increasing trend as n_0 increases. In particular, pNP-LDA is the most conservative one with the largest violation rate being 0.007 in Example 1a and 0.028 in Example 1b. In contrast, eLDA exhibits a much more accurate targeting at the violation rates, with all the observed violation rates around $\delta = 0.1$. Theorem 1 indicates that the type I error upper bound of eLDA might be violated with probability at most $\delta + C_1 n^{-1/2+\varepsilon}$. As the sample size increases, this quantity gets closer to δ. The control of violation rates for feLDA is not desirable for small n_0. However, we observe a decreasing pattern as n_0 increases, which agrees with Corollary 1. When $n_0 = 1000$, for Example 1a, the violation rate of feLDA reaches the targeted level $\delta = 0.1$. In the case of the vanilla plug-in classifier, the simulation results consistently demonstrate type I error violation rates that far exceed the target across various values of n_0. This outcome is not unexpected, as the threshold in $\hat{\phi}^*_\alpha(x)$ is not a tight high-probability upper bound for the threshold in the intermediate oracle $\tilde{\phi}^*_\alpha(x)$. Recall that $\tilde{\phi}^*_\alpha(x)$ uses the same scoring function as $\hat{\phi}^*_\alpha(x)$, but has a type I error exactly equal to α. .

Figure 1: Examples 1a, type I and type II errors for competing methods with increasing balanced sample sizes.



(a) Example 1a, type I error

(b) Example 1a, type II error

The common setting shared by Examples 1c, 1c' and 1c* includes balanced and fixed sample sizes, and increasing dimension p. Similarly, in the main text, we only present performance of Example 1c in Figure 2 and leave the comparison across Examples 1c, 1c' and 1c* to Appendix Figure 5. First, we observe from Figure 2 that both eLDA and feLDA

[2] Strictly speaking, the observed violation rate on type I error is only an approximation to the real violation rate. The approximation is two-fold: 1). in each repetition of an experiment, the population type I error is approximated by the empirical type I error on a large test set; 2). the violation rate should be calculated based on infinite repetitions of the experiment, but we only calculate it based on a finite number of repetitions. However, such approximation is unavoidable in numerical studies.

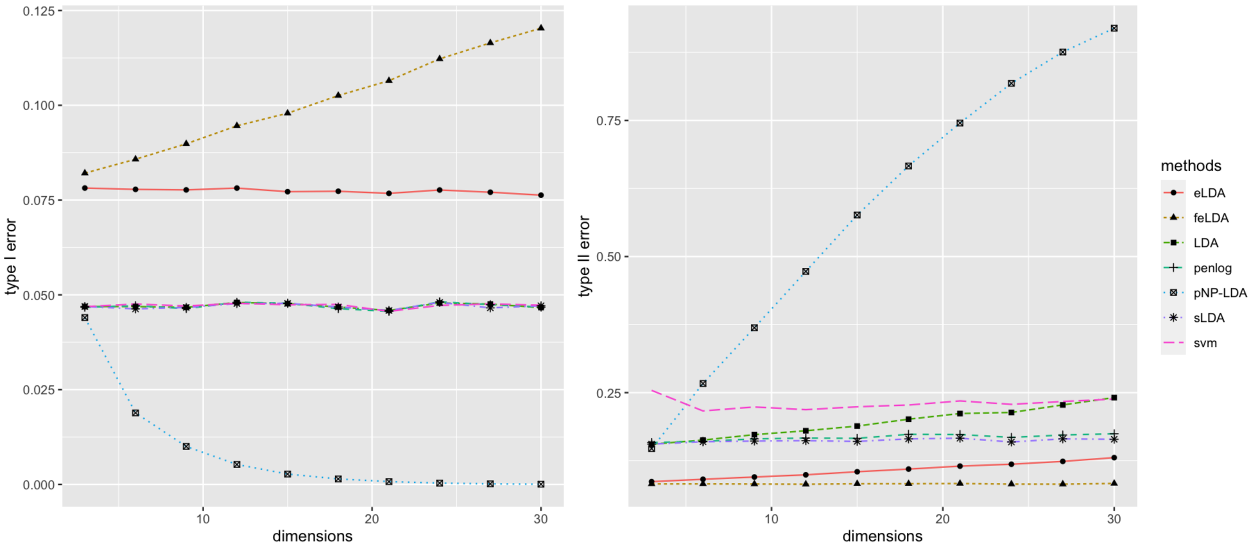
Table 2: Examples 1a and 1b, violation rates over different n_0 and methods.

	Methods	$n_0 = 20$	70	120	170	220	270	320	370	500	1000
Example 1a	NP-lda	NA	.016	.047	.062	.071	.087	.074	.074	.078	.080
	NP-slda	NA	.016	.046	.062	.071	.086	.074	.074	.077	.079
	NP-penlog	NA	.018	.050	.064	.075	.096	.071	.071	.084	.078
	NP-svm	NA	.020	.045	.064	.077	.084	.068	.064	.082	.084
	pNP-lda	.000	.000	.004	.004	.002	.001	.001	.002	.002	.007
	elda	.091	.084	.108	.105	.103	.104	.104	.100	.101	.081
	felda	.220	.144	.145	.138	.134	.141	.141	.126	.121	.100
	vanilla-plugin	.654	.568	.559	.553	.533	.535	.533	.530	.529	.547
Example 1b	NP-lda	NA	.017	.043	.055	.072	.090	.078	.069	.078	.078
	NP-slda	NA	.017	.043	.056	.072	.090	.075	.069	.077	.078
	NP-penlog	NA	.016	.047	.063	.076	.091	.075	.072	.084	.074
	NP-svm	NA	.022	.058	.066	.072	.089	.070	.065	.082	.075
	pNP-lda	.028	.015	.012	.010	.005	.005	.003	.005	.002	.000
	elda	.083	.087	.090	.095	.095	.090	.099	.102	.101	.091
	felda	.138	.122	.112	.118	.122	.121	.121	.122	.121	.112

dominate existing methods in terms of type II errors. Nevertheless, Example 1c shows that when p gets to 20 and beyond, type I error of feLDA is no longer bounded by $\alpha = 0.1$. Changing the violation rate δ from 0.1 to 0.05 and further to 0.01 hinders the growth of type I error of feLDA as p increases, but does not solve the problem ultimately as illustrated in Appendix Figure 5 panel (c) and (e). This is due to the construction of feLDA which is specifically designed for small p; when p gets large, eLDA outperforms feLDA. Therefore, considering the performance across different p's, eLDA performs the best among the seven methods. Second, as dimension p increases, all of the type II errors slightly increase or remain stable as expected, except for that of pNP-LDA. This is due to a technical bound in the construction of the threshold of pNP-LDA, which becomes loose when p is large.

Table 3 presents the violation rates from Examples 1c, 1c', and 1c*. Similar to what we have observed earlier, the five existing NP classifiers are relatively conservative and the observed violation rates of eLDA are mostly around the targeted δ in all the three sub-examples, while that of feLDA goes beyond the targeted δ as p increases. When we decrease δ from 0.1 to 0.05 and further to 0.01, we have the following two observations: 1) the violation rates of the four NP umbrella algorithm based classifiers NP-LDA, NP-sLDA, NP-penlog and NP-svm stay the same in Examples 1c and 1c'. The violation rates decrease as we move to Example 1c*. This is due to the discrete combinatorial construction of the thresholds in umbrella algorithms and thus the observed violation rates present discrete changes in terms of δ. In other words, not necessarily small changes in δ will lead to a change in the constructed classifier and the observed violation rates. For example, for NP umbrella algorithm based methods, the number of left-out class 0 observations is 63, and the threshold is constructed as the k^*-th order statistics of the classification scores of the left-out class 0 sample, where $k^* = \min\{k \in \{1, \cdots, 63\} : \nu(k) < \delta\}$, and $\nu(k) = \sum_{j=k}^{63} \binom{63}{j} (1-\alpha)^j \alpha^{63-j}$. Plugging in $\alpha = 0.1$, we could easily calculate that $k^* = 61$ for both $\delta = 0.1$ and $\delta = 0.05$, since $\nu(61) = \sum_{j=61}^{63} \binom{63}{j} (1-0.1)^j 0.1^{63-j} = 0.042$ and $\nu(60) = 0.113$. Furthermore, for $\delta = 0.01$, the threshold changes as k^* changes, since $0.042 > 0.01$; 2) pNP-LDA, eLDA, and feLDA have the parametric construction of the threshold and the observed violation rates of these methods respond to changes in δ more smoothly. Nevertheless, pNP-LDA is overly conservative, with the observed violation rate almost all 0.

Figure 2: Examples 1c, type I and type II errors for competing methods with increasing dimension p, $\delta = 0.1$.



(a) Example 1c, type I error

(b) Example 1c, type II error

Table 3: Examples 1c, 1c' and 1c*, violation rates over different p and methods.

	Methods	$p = 3$	6	9	12	15	18	21	24	27	30
Example 1c $\delta = 0.1$	NP-lda	.044	.039	.039	.045	.058	.046	.035	.049	.044	.048
	NP-slda	.045	.033	.037	.050	.047	.043	.034	.045	.038	.041
	NP-penlog	.037	.042	.035	.056	.050	.044	.031	.049	.043	.041
	NP-svm	.041	.040	.041	.044	.041	.042	.043	.039	.035	.048
	pNP-lda	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000
	elda	.105	.091	.084	.107	.104	.079	.105	.099	.082	.082
	felda	.147	.206	.274	.362	.435	.548	.597	.712	.790	.817
Example 1c' $\delta = 0.05$	NP-lda	.044	.039	.039	.045	.058	.046	.035	.049	.044	.048
	NP-slda	.045	.033	.037	.050	.047	.043	.034	.045	.038	.041
	NP-penlog	.037	.042	.035	.056	.050	.044	.031	.049	.043	.041
	NP-svm	.041	.040	.041	.044	.041	.042	.043	.039	.035	.048
	pNP-lda	.001	.000	.000	.000	.000	.000	.000	.000	.000	.000
	elda	.061	.049	.043	.052	.046	.042	.057	.054	.044	.044
	felda	.087	.115	.161	.260	.431	.410	.472	.599	.679	.732
Example 1c* $\delta = 0.01$	NP-lda	.001	.001	.000	.004	.002	.000	.000	.000	.002	.001
	NP-slda	.001	.001	.001	.003	.004	.000	.002	.000	.000	.001
	NP-penlog	.001	.001	.000	.003	.004	.000	.001	.000	.001	.001
	NP-svm	.000	.002	.001	.002	.000	.002	.000	.000	.001	.001
	pNP-lda	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
	elda	.014	.007	.008	.009	.003	.011	.011	.016	.010	.010
	felda	.025	.032	.053	.100	.146	.188	.259	.361	.436	.530

Examples 1d and 1d' also demonstrate the performances when dimension p increases, but with unequal class sizes. We omit the details in the main due to similar messages, and refer interested readers to Appendix Figure 6.

Example 2 *The data are generated from an LDA model with common covariance matrix Σ, where Σ is set to be an AR(1) covariance matrix with $\Sigma_{ij} = 0.5^{|i-j|}$ for all i and j.*

$\boldsymbol{\beta}^{Bayes} = \Sigma^{-1}\boldsymbol{\mu}_d = C_p \cdot \mathbf{1}_p^\top$, $\boldsymbol{\mu}^0 = \mathbf{0}_p$. Here, C_p is a constant depending on p, such that the NP oracle classifier always has type II error 0.236 for any choice of p when $\alpha = 0.1$. We set $\pi_0 = \pi_1 = 0.5$ and $\alpha = \delta = 0.1$.

(2a) $n_0 = n_1 = 125$, varying $p \in \{3, 6, 9, 12, 15, 18, 21, 24, 27, 30\}$

(2b) $n_0 = 125, n_1 = 500$, varying $p \in \{3, 6, 9, 12, 15, 18, 21, 24, 27, 30\}$

Examples 2a and 2b are similar to Examples 1c and 1d, but their oracle projection direction $\boldsymbol{\beta}^{\text{Bayes}}$ is not sparse. Appendix Figure 7 summarizes the results on type I and type II errors. The delivered messages are similar to those of Examples 1c and 1d: 1) while `eLDA` enjoys controlled type I errors under $\alpha = 0.1$ for all p in both Examples 2a and 2b, the type I errors of `feLDA` deteriorate above the target for large p; 2) `eLDA` and `feLDA` dominate all other competing methods in terms of type II errors. Observed violation rates from Examples 2a and 2b present similar messages as in Examples 1c and 1d, so we omit the table for those results.

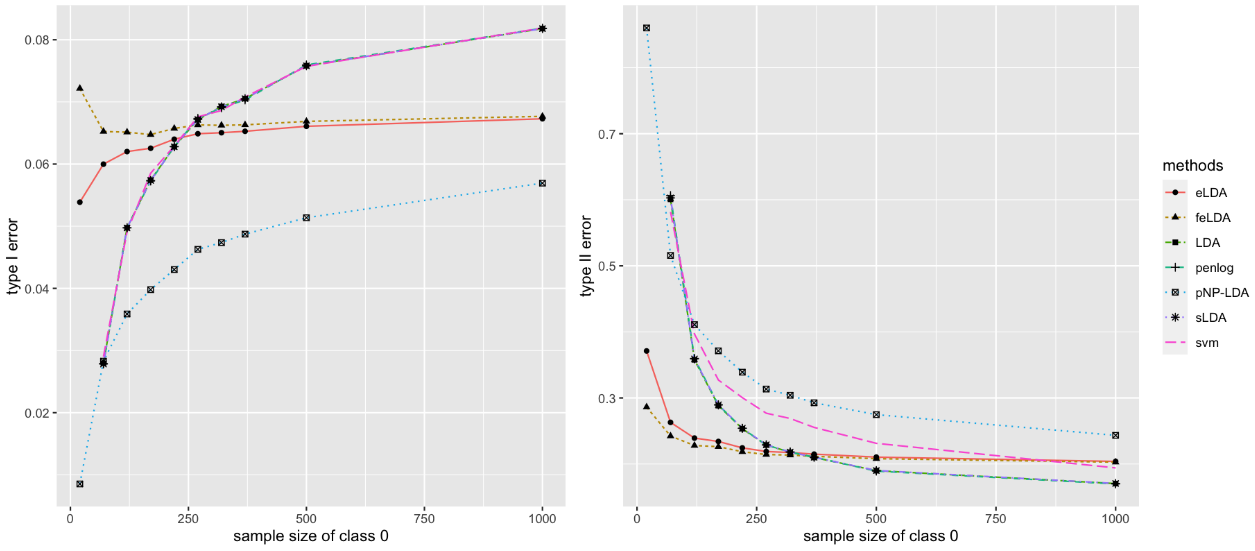
Example 3 *The data are generated from multivariate t-distributions with degrees of freedom 4. Two classes share a common covariance matrix Σ, where $\Sigma_{ij} = 0.5^{|i-j|}$ for all i and j. $\boldsymbol{\mu}^0 = \mathbf{0}_p$ and $\boldsymbol{\mu}^1$ is defined by $\Sigma^{-1}\boldsymbol{\mu}_d = 1.2 \times \mathbf{1}_p^\top$ with $p = 3$. We set $\pi_0 = \pi_1 = 0.5$, $\alpha = \delta = 0.1$. In terms of dimensions, we let the balanced sample sizes n_0 and n_1 grow together as $n_0 = n_1 \in \{20, 70, 120, 170, 220, 270, 320, 370, 500, 1000\}$. Type I and type II errors are evaluated on a test set with 30,000 observations from each class, and then we report the average over the 1,000 repetitions.*

Example 3 helps provide a broader understanding of the newly proposed classifiers under non-Gaussian distributions. Figure 3 depicts type I and type II errors, and Table 4 summarizes the observed violation rates. We have two observations as follows: 1) among `pNP-lda`, `elda` and `felda`, which are implementable for all sample sizes, `elda` and `felda` clearly dominate `pNP-lda`. `elda` and `felda` have the type I error bounded under α and enjoy much smaller type II errors comparing to `pNP-lda`; 2) comparing `elda` and `felda` with other umbrella algorithm based NP classifiers, we observe that when sample size of class 0 is very small (in the current setting, less than 220), the umbrella algorithm based classifiers either cannot be implemented ($n_0 = 20$) or have much worse type II errors than `elda` and `felda`. As the sample size further increases, the performances of most umbrella algorithm based classifiers begin to catch up and eventually outperform `elda` and `felda`. We believe this phenomenon is due to the fine calibration of the LDA model in the development of `elda` and `felda`, which leads to conservative results in heavy-tail distribution settings. On the other hand, the nonparametric NP umbrella algorithm does not rely on any distributional assumptions and benefit from larger sample sizes.

5.2 Real Data Analysis

In this section, we analyze five datasets in total, including two cancer datasets, Fashion MNIST, a dataset on spam email detection and the CSE-CIC-IDS2018 dataset on network intrusion. One thing to keep in mind is that, the observed violation rate in real data analysis should not be interpreted as a close proxy to the true violation rate. First, the

Figure 3: Example 3, type I and type II errors for competing methods with increasing and balanced sample sizes



(a) Example 3, Type I errors

(b) Example 3, Type II errors

Table 4: $\delta = 0.1$, $p = 3$; violation rates over different n_0 and methods.

	Methods	$n_0 = 20$	70	120	170	220	270	320	370	500	1000
Example 3	`NP-lda`	NA	.024	.067	.064	.068	.074	.073	.078	.073	.057
	`NP-slda`	NA	.024	.071	.062	.069	.078	.069	.076	.074	.056
	`NP-penlog`	NA	.021	.061	.059	.069	.077	.073	.074	.075	.058
	`NP-svm`	NA	.026	.063	.066	.066	.084	.065	.080	.081	.068
	`pNP-lda`	.000	.001	.000	.000	.000	.000	.000	.000	.000	.000
	`elda`	.075	.026	.009	.006	.003	.001	.003	.001	.000	.000
	`felda`	.191	.043	.017	.011	.004	.001	.003	.001	.001	.000

previous discussion on observed type I error violation rate for simulation in the footnote also applies to the real data studies. Moreover, in simulations, samples are generated from population many times; however, in real data analysis, the one sample we have plays the role of population for repetitive sampling. Such substitute can be particularly inaccurate when the sample size is small.

Based on the findings from the five datasets, the main insights are as follows: 1) when the class 0 sample size is insufficient to meet the sample size requirement for the NP umbrella algorithm, two viable options remain: `pNP-lda` and `eLDA`. However, `pNP-lda` tends to be excessively conservative, resulting in an observed type I error violation rate of 0 and type II error of 1. On the other hand, the newly proposed `eLDA` achieves reasonable type II errors while effectively controlling the violation rate; 2) when the covariate distributions significantly deviate from Gaussian, `eLDA` can still achieve the desired control over type I error, demonstrating its robustness to non-Gaussian distributed features; and 3) in cases where the sample size is small compared to the feature dimensionality, a pre-step for `eLDA`

Table 5: Lung cancer dataset

		pNP-LDA	eLDA
$\alpha = 0.01$ $\delta = 0.05$	type I error	.000	.003
	type II error	1	.104
	observed violation rate	0	.03

is to first select only the important features, yielding a practical way to apply `eLDA` more widely. Next, we will thoroughly analyze and closely examine the results of each dataset.

5.2.1 Lung Cancer Dataset

The first dataset is a lung cancer dataset (Gordon et al., 2002; Jin and Wang, 2016) that consists of gene expression measurements from 181 tissue samples. Among them, 31 are malignant pleural mesothelioma (MPM) samples and 150 are adenocarcinoma (ADCA) samples. As MPM is known to be highly lethal pleural malignant and rare (in contrast to ADCA which is more common), misclassifying MPM as ADCA would incur more severe consequences. Therefore, we code MPM as class 0, and ADCA as class 1. The feature dimension of this dataset is $p = 12{,}533$. First, we set $\alpha = 0.01$ and $\delta = 0.05$. Since the class 0 sample size is very small, none of the umbrella algorithm based NP classifiers are implementable. Hence, we only compare the performance of `pNP-LDA` with that of `eLDA`. We choose to omit `feLDA` here because we have found from the simulation studies that `feLDA` outperforms `eLDA` only when the dimension is extremely small (e.g., $p \leq 3$). On the other hand, since `eLDA` is designed for $p < n$ settings and `pNP-LDA` usually works poorly for large p, we first reduce the feature dimensionality to 40 by conducting two-sample t-test and selecting the 40 genes with smallest p-values. To provide a more complete story, we implemented further analysis with larger parameters ($\alpha = 0.1$ and $\delta = 0.4$) so that `NP-sLDA`, `NP-penlog`, `NP-svm` are also implementable. Those results are presented in Appendix Table 10.

The experiment is repeated 100 times and the type I and type II errors are the averages over these 100 replications. In each replication, we randomly split the full dataset (class 0 and class 1 separately) into a training set (composed of 70% of the data), and a test set (composed of 30% of the data). We train the classifiers on the training set, with the feature selection step added before implementing `eLDA` and `pNP-LDA`. Then we apply the classifiers to the test set to compute the empirical type I and type II errors. Table 5 presents results from the parameter set $\alpha = 0.01$ and $\delta = 0.05$. We observe that while both `eLDA` and `pNP-LDA` achieve type I errors smaller than the targeted $\alpha = 0.01$, `pNP-LDA` is overly conservative and has a type II error of 1. In contrast, `eLDA` provides a more reasonable type II error of 0.104, and the observed violation rate is 0.03 (< 0.05).

5.2.2 A Microarray Dataset of 11 Tumor Types (Su et al., 2001)

The second dataset was originally studied in (Su et al., 2001). It contains microarray data from 11 different tumor types, including 27 serous papillary ovarian adenocarcinomas, 8 bladder/ureter carcinomas, 26 infiltrating ductal breast adenocarcinomas, 23 colorectal adenocarcinomas, 12 gastroesophageal adenocarcinomas, 11 clear cell carcinomas of the kidney, 7 hepatocellular carcinomas, 26 prostate adenocarcinomas, 6 pancreatic adenocar-

Table 6: Cancer dataset in (Su et al., 2001)

		pNP-LDA	eLDA
$\alpha = 0.01$ $\delta = 0.05$	type I error	.000	.008
	type II error	1	.437
	observed violation rate	0	.15

cinomas, 14 lung adenocarcinomas carcinomas, and 14 lung squamous carcinomas. In more recent studies (Jin and Wang, 2016; Yousefi et al., 2010), the 11 different tumor cell types were aggregated into two classes, where class 0 contains bladder/ureter, breast, colorectal and prostate tumor cells, and class 1 contains the remaining groups. We follow (Yousefi et al., 2010) in determining the binary class labels, and we work on the modified dataset with $n_0 = 83$, $n_1 = 91$ and $p = 12{,}533$.

We repeat the data processing procedure as in the lung cancer dataset, and report results from the parameter set $\alpha = 0.01$ and $\delta = 0.05$ in Table 6. While the sample size is too small for other umbrella algorithm based NP classifiers to work, the advantage of `eLDA` over `pNP-LDA` is obvious.

5.2.3 Fashion MNIST

Fashion MNIST is a widely-used imaging dataset for benchmarking machine learning algorithms. It contains 60,000 training data and 10,000 testing data from ten different fashion categories, including T-shirt/top, Trousers, Sneakers, Bags, and others. Despite numerous algorithms achieving near-perfect results over the entire dataset, we focus on a regime with a small subset of the training data to demonstrate the benefits of utilizing a parametric model in the face of limited training data.

Specifically, we subsample 10% of the training data in each repetition while keeping a fixed testing set with 10,000 images. We report two types of errors and the violation rate of type I error over 100 repetitions. Given that the eLDA algorithm handles binary classifications only, we conduct two experiments with different grouping strategies, each time selecting one subcategory as class 0 and combining the remaining nine categories as class 1, resulting in highly unbalanced sample sizes. The remainder of the data processing is the same as that used in cancer datasets.

From Table 7, we have the following observations. First, when the sample sizes are small, even if the sample size requirement is satisfied, umbrella algorithms, i.e., `NP-slda`, `NP-penlog` and `NP-svm` may fail to control the violation rate. Second, while both `eLDA` and `pNP-lda` conservatively guarantee the violation rate, `eLDA` achieves a much smaller type II error (0.488 versus 1 and 0.438 versus 1), in both of the experimental settings.

5.2.4 Spam Email

We analyzed another dataset containing 4,601 spam and non-spam emails, each with 57 word frequency-related attributes. To avoid labeling good emails as spam which is the more severe type of error, we labeled non-spams as class 0 and spam as class 1.

For the NP umbrella algorithms, we retained all features, while for `pNP-lda` and `eLDA`, we performed feature selection. To demonstrate the functionality of `eLDA`, we selected the top 20 features with the largest p-values from Shapiro-Wilk tests, since `eLDA` is based on

Table 7: Fashion MNIST with $\alpha = 0.1$ and $\delta = 0.1$

		`NP-slda`	`NP-penlog`	`NP-svm`	`pNP-lda`	`eLDA`
{T-shirt/top}	`type I error`	.087	.084	.082	.000	.052
v.s.	`type II error`	.346	.308	.309	1.000	.488
{Others}	`violation rate`	.21	.17	.09	.00	.00
{Sandal}	`type I error`	.089	.087	.086	.000	.007
v.s.	`type II error`	.243	.237	.243	1.000	.438
{Others}	`violation rate`	.28	.21	.20	.00	.00

an LDA model. Nevertheless, this was a fair comparison since we used the full model for the NP umbrella algorithms, and reported better results for these algorithms comparing to using only the top 20 features. Moreover, the top 20 features selected had p-values ranging from 10^{-55} to 10^{-84}, indicating that they were still far from normally distributed, and thus our `eLDA` adapted well to non-Gaussian distributions.

We report results from two scenarios with $\alpha = 0.05$ and $\delta = 0.01$. In the first scenario, we randomly selected 10% of the dataset as training data, providing a sufficiently large sample size for all methods to work. In the second scenario, we randomly selected 5% of the dataset as training data, creating a sample size too small for the NP umbrella algorithms to work. Therefore, we only compared the performance of `eLDA` and `pNP-lda`.

We observe from Table 8 that, under both of the cases, violation rates are constantly under control. When the sample size is large enough, `eLDA` achieves a smaller type II error than both `NP-slda`, which is constructed based on the same LDA assumption, and `NP-penlog`. The type II error of `eLDA` is slightly larger than `NP-svm`. In the regime where the sample size is too small for the NP umbrella algorithms to work, the advantage of `eLDA` is apparent, with a type II error of 0.581 comparing with 1 which is realized by `pNP-lda`. When we decrease the training sample size from 10% to 5% of the whole available dataset, the type II error of `eLDA` increases, but only mildly. This Spam Email example suggested us a practical way to use `eLDA`, that is when the sample size is small, instead of using all the features for `eLDA`, we may screen out the ones that are most Gaussian-like, and performance may even be better than using the full model for the NP umbrella algorithms.

Table 8: Spambase Dataset with $\alpha = 0.05$ and $\delta = 0.01$

		`NP-slda`	`NP-penlog`	`NP-svm`	`pNP-lda`	`eLDA`
training set:	`type I error`	.012	.013	.014	.000	.027
10% of the	`type II error`	.768	.707	.509	1.000	.550
dataset	`violation rate`	.00	.00	.01	.00	.00
training set	`type I error`	NA	NA	NA	.000	.028
5% of the	`type II error`	NA	NA	NA	1.000	.581
dataset	`violation rate`	NA	NA	NA	.00	.00

5.2.5 CSE-CIC-IDS2018

We consider the popular network intrusion classification problem and apply the NP classifiers to the CSE-CIC-IDS2018 dataset (Sharafaldin et al., 2018). It was a collaborative project between the Communications Security Establishment (CSE) and the Canadian Institute for Cybersecurity (CIC). The original dataset contains seven different malicious types of attack scenarios including Brute-force, Heartbleed, Botnet and etc. Since our method focuses on binary classification, we only consider one type of the malicious attack scenarios, the FTP-BruteForce attacks, and define it as class 0. We define benign attacks as class 1. Then we have 193,360 observations in class 0, 667,626 observations in class 1 with 79 features in total. As a benchmark dataset, the overall accuracy has been very high; nevertheless, we would like to demonstrate the performance of our method for such type of data when the available sample size is relatively small. Therefore, similar to the Spambase Dataset in Section 5.2.4, we randomly subsample a small proportion of the whole dataset to be our training data and evaluate the performance on the rest of the data. We conduct analysis in the following three scenarios and discuss the findings: 1) $\alpha = 0.05$ and $\delta = 0.2$, 0.04% of the dataset as training (all methods work); 2) $\alpha = 0.05$ and $\delta = 0.2$, 0.02% of the dataset as training (only `pNP-lda` and `eLDA` work); 3) $\alpha = 0.05$ and $\delta = 0.1$, 0.04% of the dataset as training (only `pNP-lda` and `eLDA` work).

Comparing the first two scenarios, we observe from Table 9 that, when the sample size is large enough, while all the methods achieve desirable type I error control and the targeted violation rate, it is beneficial to apply the non-model-based NP classifiers with the umbrella algorithms as they achieve smaller type II errors. Nevertheless, when we decrease the training data to include only 0.02% of the dataset, `NP-slda`, `NP-penlog` and `NP-svm` fail to work due to the minimum sample size requirement. In this regime, the advantage of `eLDA` over `pNP-lda` is tremendous, with `eLDA` achieving a type II error of 0.022, while `pNP-lda` achieved a type II error of 1. When we select 0.04% of the dataset as training data, together with the parameter sets as $\alpha = 0.05$ and $\delta = 0.1$, `NP-slda`, `NP-penlog`, `NP-svm` cannot be applied either. In this case, `eLDA` achieved a type I error of 0.010, a type II error of 0.019 and a violation rate of 0.08, which is desirable and much superior than the only applicable alternative `pNP-lda`.

Table 9: CSE-CIC-IDS Dataset with $\alpha = 0.05$ and $\delta = 0.2$

		`NP-slda`	`NP-penlog`	`NP-svm`	`pNP-lda`	`eLDA`
training set: 0.04% of the dataset	`type I error`	.025	.025	.025	.000	.011
	`type II error`	0	0	0	1.000	.018
	`violation rate`	.12	.12	.12	.00	.10
training set 0.02% of the dataset	`type I error`	NA	NA	NA	.000	.012
	`type II error`	NA	NA	NA	1.000	.022
	`violation rate`	NA	NA	NA	.00	.1

6. Discussion

Our current work initiates the investigations on non-splitting strategies under the NP paradigm. With the explicit LDA model assumption, we can use the model characteristics and distributional properties of the estimates to achieve high-probability type I error control, circumventing the minimum class 0 sample size requirement as in the NP umbrella algorithm which relies on an order statistics approach.

For future works, we can work in settings where p is larger than n by selecting features via various marginal screening methods (Fan and Song, 2010; Li et al., 2012) and/or may add structural assumptions to the LDA model. To accommodate diverse applications, one might also construct classifiers based on more complicated models, such as the quadratic discriminant analysis (QDA) model (Fan et al., 2015; Li and Shao, 2015; Yang and Cheng, 2018; Pan and Mai, 2020; Wang et al., 2021; Cai and Zhang, 2021).

Acknowledgment

The authors would like to thank the Action Editor and two anonymous referees for many constructive comments which greatly improved the paper. This work was partially supported by National Science Foundation grant DMS-2113500, Hong Kong RGC grants 26305120 and 16303922.

Appendix A. Further remark on Assumption 1

Previously, margin assumption and detection condition were assumed in Tong (2013) and subsequent works (Zhao et al., 2016; Tong et al., 2020) for an NP classifier to achieve a diminishing excess type II error. Concretely, write the level-α NP oracle as $\mathbb{1}(f_1(\mathbf{x})/f_0(\mathbf{x}) > C^*_\alpha)$, where f_1 and f_0 are class-conditional densities of the features, then the margin assumption assumes that

$$\mathbb{P}(|f_1(\mathbf{x})/f_0(\mathbf{x}) - C^*_\alpha| \leq \delta | Y = 0) \leq C_0 \delta^{\bar{\gamma}} \,,$$

for any $\delta > 0$ and some positive constant $\bar{\gamma}$ and C_0. This is a low-noise condition around the oracle decision boundary that has roots in (Polonik, 1995; Mammen and Tsybakov, 1999). On the other hand, the detection condition, which was coined in Tong (2013) and refined in Zhao et al. (2016), requires a lower bound:

$$\mathbb{P}(C^*_\alpha \leq f_1(\mathbf{x})/f_0(\mathbf{x}) \leq C^*_\alpha + \delta | Y = 0) \geq C_1 \delta^{\underline{\gamma}} \,,$$

for small δ and some positive constant $\underline{\gamma}$. In fact, $\delta^{\underline{\gamma}}$ can be generalized to $u(\delta)$, where $u(\cdot)$ is any increasing function on R^+ that might be (f_0, f_1)-dependent and $\lim_{\delta \to 0^+} u(\delta) = 0$. The necessity of the detection condition under general models for achieving a diminishing excess type II error was also demonstrated in Zhao et al. (2016) by showing a counterexample that has fixed f_1 and f_0, i.e., when p does not grow with n. Note that although the feature dimension p considered in Zhao et al. (2016) and Tong et al. (2020) can grow with n, both impose sparsity assumptions, and the "effective" dimensionality s has the property that $s/n \to 0$. Hence previously, there were no theoretical results regarding the excess type II error when the effective feature dimensionality and the sample size are comparable.

Under Assumption 1, the marginal assumption and detection condition hold automatically. To see this, recall the level-α NP oracle classifier defined in (2.1), we can directly derive that for any $\delta > 0$,

$$\begin{aligned}
&\mathbb{P}(C^*_\alpha \le f_1(\mathbf{x})/f_0(\mathbf{x}) \le C^*_\alpha + \delta | Y = 0) \\
&= \mathbb{P}(F \le (\Sigma^{-1}\boldsymbol{\mu}_d)^\top \mathbf{x} \le F + \delta | Y = 0) \\
&= \mathbb{P}\big(F - \boldsymbol{\mu}_d^\top \Sigma^{-1}\boldsymbol{\mu}^0 \le (\Sigma^{-1}\boldsymbol{\mu}_d)^\top(\mathbf{x} - \boldsymbol{\mu}^0) \le F - \boldsymbol{\mu}_d^\top\Sigma^{-1}\boldsymbol{\mu}^0 + \delta | Y = 0\big) \\
&= \mathbb{P}\Big(\frac{F - \boldsymbol{\mu}_d^\top\Sigma^{-1}\boldsymbol{\mu}^0}{\sqrt{\Delta_d}} \le \mathcal{N}(0,1) \le \frac{F - \boldsymbol{\mu}_d^\top\Sigma^{-1}\boldsymbol{\mu}^0 + \delta}{\sqrt{\Delta_d}}\Big),
\end{aligned}$$

with the shorthand notation $F := \sqrt{\Delta_d}\,\Phi^{-1}(1-\alpha) + \boldsymbol{\mu}_d^\top\Sigma^{-1}\boldsymbol{\mu}^0$. The RHS above can be further simplified to get

$$\mathbb{P}(F \le (\Sigma^{-1}\boldsymbol{\mu}_d)^\top\mathbf{x} \le F + \delta | Y = 0) = \Phi\Big(\Phi^{-1}(1-\alpha) + \delta/\sqrt{\Delta_d}\Big) - (1-\alpha).$$

Thereby, using mean value theorem, we simply bound the above probability from above and below as

$$\begin{aligned}
&\mathbb{P}(F \le (\Sigma^{-1}\boldsymbol{\mu}_d)^\top\mathbf{x} \le F + \delta | Y = 0) \le \frac{1}{\sqrt{2\pi}}\exp\big(-\frac{1}{2}\Phi_\alpha^2\big)\frac{\delta}{\sqrt{\Delta_d}}, \\
&\mathbb{P}(F \le (\Sigma^{-1}\boldsymbol{\mu}_d)^\top\mathbf{x} \le F + \delta | Y = 0) \ge \frac{1}{\sqrt{2\pi}}\exp\Big(-\frac{1}{2}\Big(\Phi_\alpha + \frac{\delta}{\sqrt{\Delta_d}}\Big)^2\Big)\frac{\delta}{\sqrt{\Delta_d}}.
\end{aligned}$$

where we recall $\Phi_\alpha = \Phi^{-1}(1-\alpha)$. A similar upper bound can also be derived for $\mathbb{P}(F - \delta \le (\Sigma^{-1}\boldsymbol{\mu}_d)^\top\mathbf{x} \le F | Y = 0)$. These coincide with the aforementioned marginal assumption and detection condition.

Appendix B. Proofs of Lemma 1 and Corollary 1

We first show the proof of Lemma 1 below.

Proof [Proof of Lemma 1] The statement (i) is easy to obtain by the definition of $\tilde{\phi}^*_\alpha(\cdot)$ in (3.1) and the definition of the type I error. Specifically,

$$\begin{aligned}
R_0(\tilde{\phi}^*_\alpha) &= \mathbb{P}\Big(\widehat{A}^\top(\mathbf{x} - \boldsymbol{\mu}^0) > \sqrt{\widehat{A}^\top\Sigma\widehat{A}}\,\Phi^{-1}(1-\alpha)\Big|\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}^0, \Sigma)\Big) \\
&= 1 - \Phi(\Phi^{-1}(1-\alpha)) = \alpha.
\end{aligned}$$

Next, we establish statement (ii). By definition, we have

$$\begin{aligned}
R_1(\tilde{\phi}^*_\alpha) &= \mathbb{P}\Big(\widehat{A}^\top(\mathbf{x} - \boldsymbol{\mu}^1) \le \sqrt{\widehat{A}^\top\Sigma\widehat{A}}\,\Phi^{-1}(1-\alpha) - \widehat{A}^\top\boldsymbol{\mu}_d\Big|\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}^1, \Sigma)\Big) \\
&= \Phi\Big(\Phi^{-1}(1-\alpha) - \frac{\widehat{A}^\top\boldsymbol{\mu}_d}{\sqrt{\widehat{A}^\top\Sigma\widehat{A}}}\Big), \\
R_1(\phi^*_\alpha) &= \mathbb{P}\Big((\Sigma^{-1}\boldsymbol{\mu}_d)^\top\mathbf{x} < \sqrt{\Delta_d}\,\Phi^{-1}(1-\alpha) + \boldsymbol{\mu}_d^\top\Sigma^{-1}\boldsymbol{\mu}^0\Big|\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}^1, \Sigma)\Big) \\
&= \Phi\Big(\Phi^{-1}(1-\alpha) - \sqrt{\Delta_d}\Big). \qquad \text{(B.1)}
\end{aligned}$$

Lemma 3 and some elementary calculations lead to the conclusion: for any $\varepsilon \in (0, 1/2)$ and $D > 0$, when $n > n(\epsilon, D)$, with probability at least $1 - n^{-D}$ we have,

$$\Delta_d^{1/2} > \frac{\widehat{A}^\top \boldsymbol{\mu}_d}{\sqrt{\widehat{A}^\top \Sigma \widehat{A}}} = \Delta_d^{1/2} + O\big(r\Delta_d^{1/2}\big) + O\big(n^{-\frac{1}{2}+\varepsilon}(1 + \Delta_d^{1/2})\big)\,.$$

Moreover, it is straightforward to check

$$\exp\Big(-\frac{1}{2}\Big(\Phi^{-1}(1-\alpha) - \Delta_d^{1/2}\Big)^2\Big) \asymp \exp\Big(-\frac{c\Delta_d}{2}\Big)\,.$$

Thus, we conclude that there exists some fixed constant C which may depend on c_0, c_1, c_2 and α such that for any $\varepsilon \in (0, 1/2)$ and $D > 0$, when $n \geq n(\varepsilon, D)$, with probability at least $1 - n^{-D}$, we have

$$R_1(\tilde{\phi}^*_\alpha) - R_1(\phi^*_\alpha) \leq C\big(r + n^{-\frac{1}{2}+\varepsilon}\big)\Delta_d^{1/2} \exp\Big(-\frac{c\Delta_d}{2}\Big)\,.$$

Let $D = 1$, we thus finished our proof. ■

At the end of this section, we sketch the proof of Corollary 1.

Proof [Proof of Corollary 1] By the definition of $\widetilde{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0)$ and $\widetilde{C}^p_\alpha$ in (3.7), (3.8), under the setting of $p = O(1)$, we observe that

$$\begin{aligned}\widetilde{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0) &= \widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0) + O_\prec\big(n^{-1}\Delta_d^{\frac{1}{2}}\big)\,,\\ \widetilde{C}^p_\alpha &= \widehat{C}^p_\alpha + O_\prec\big(n^{-1}\Delta_d^{\frac{1}{2}}\big)\,.\end{aligned}$$

Then, similarly to the proof of Theorem 1, with the aid of Remark 3 and Remark 4, we conclude the results in the same manner; hence we omit the details.

■

Appendix C. Proofs for Section 4.1

C.1 Proof of Lemma 2

Part (i) is obvious from Definition 1. For any fixed $\varrho > 0$, we have

$$\begin{aligned}|\mathbb{E}\mathsf{X}_1| &\leq \mathbb{E}|\mathsf{X}_1 \mathbb{1}(|\mathsf{X}_1| \leq N^\varrho \Phi)| + \mathbb{E}|\mathsf{X}_1 \mathbb{1}(|\mathsf{X}_1| \geq N^\varrho \Phi)|\\ &\leq N^\varrho \Phi + N^C \mathbb{P}(|\mathsf{X}_1| \geq N^\varrho \Phi) = O(N^\varrho \Phi)\end{aligned}$$

for for sufficiently large $N \geq N_0(\varrho)$. This proves part (ii).

C.2 Proof of Proposition 1

Define

$$\mathcal{D} \equiv \mathcal{D}(\tau) := \{z \in \mathbb{C}^+ : -\frac{\lambda_-}{2} < \Re z < \frac{\lambda_-}{2}, 0 < \Im z \leq \tau^{-1}\}\,. \tag{C.1}$$

All the estimates in Proposition 1 can be separately shown for the case of $p > n^\epsilon$ for some fixed small $\epsilon > 0$ and the case of $p < n^\epsilon$. We first show all the estimates hold for the case $l = 0$ and then proceed to the case of $l \geq 1$.

- For the case of $l = 0$.

In the regime that $p \geq n^\epsilon$ for some fixed small $\epsilon > 0$, (4.5) can be derived from the entrywise local Marchenko-Pastur law for extended spectral domain in Theorem 4.1 of (Bloemendal et al., 2014). We emphasize that originally in (Bloemendal et al., 2014) the results are not provided for extended spectral domain one only need to adapt the arguments in Proposition 3.8 of (Bloemendal et al., 2016) to extend the results.

The estimates of (4.7) can be obtained by the rigidity estimates of eigenvalues in (Bloemendal et al., 2014, Theorem 2.10). We remark that we get the improved version in the second estimate of (4.8) due to the trivial bound $z = O(1)$, for $z \in \mathcal{D}^0$, while for $z \in \mathcal{D}$, we crudely bound $|z|$ by $r^{-\frac{1}{2}}$. For (4.6), by noticing that $X^\top \mathcal{G}_1 = \mathcal{G}_2 X$, one only needs to show the first estimate of (4.6). Using singular value decomposition (SVD) of X, i,e., $X = U^\top(\Lambda^{\frac{1}{2}}, 0)V$, where the diagonal matrix $\Lambda^{\frac{1}{2}}$ collects the singular values of X in a descending order, we arrive at

$$\big(X^\top \mathcal{G}_1(z)\big)_{i'i} = V_{i'}^\top \begin{pmatrix} \Lambda^{\frac{1}{2}}(\Lambda - z)^{-1} \\ \mathbf{0} \end{pmatrix} U_i, \qquad \Lambda := \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$$

and U_i, $V_{i'}$ are independent and uniformly distributed on $\mathbf{S}^{p-1}$ and $\mathbf{S}^{n-1}$, respectively, thanks to the fact that X is a GOE matrix. Here we abbreviate $\lambda_i(H)$ by λ_i. Then we can further write

$$\begin{aligned}\big(X^\top \mathcal{G}_1(z)\big)_{i'i} &\overset{\mathrm{d}}{=} \sum_{i=1}^{p} g_i \tilde{g}_i \frac{\sqrt{\lambda_i}}{\lambda_i - z} \frac{1}{\|\mathbf{g}\| \|\tilde{\mathbf{g}}\|} \\ &= \sum_{i=1}^{p} g_i \tilde{g}_i \frac{\sqrt{\lambda_i}}{\lambda_i - z}\Big(1 - \frac{\|\mathbf{g}\|^2 - 1}{2} + O_\prec(n^{-1})\Big)\Big(1 - \frac{\|\tilde{\mathbf{g}}\|^2 - 1}{2} + O_\prec(p^{-1})\Big),\end{aligned} \tag{C.2}$$

where $\mathbf{g} := (g_1, \cdots, g_p) \sim \mathcal{N}(0, \frac{1}{p} I_p)$, $\tilde{\mathbf{g}} := (\tilde{g}_1, \cdots, \tilde{g}_n) \sim \mathcal{N}(0, \frac{1}{n} I_n)$ and they are independent. The leading term on the RHS of (C.2) is $\sum_{i=1}^p g_i \tilde{g}_i \frac{\sqrt{\lambda_i}}{\lambda_i - z}$. By the rigidity of eigenvalues, we easily get that $\sqrt{\lambda_i}/(\lambda_i - z) \asymp r^{\frac{1}{4}}$ uniformly for $z \in \mathcal{D}$ with high probability. Further applying the randomness of g_i's and $\tilde{g}_i$'s, it is easy to conclude the first estimate in (4.6). The second estimate with the extension in (4.8) holds naturally from $X^\top \mathcal{G}_1 = \mathcal{G}_2 X$ and the facts that $|z| \leq r^{-\frac{1}{2}}$ for $z \in \mathcal{D}$, $|z| = O(1)$ for $z \in \mathcal{D}^0$.

In the regime that $p < n^\epsilon$ for sufficiently small ϵ. We first write

$$XX^\top = r^{-\frac{1}{2}} I_p + G,$$

where G is a p by p matrix defined entrywise by $G_{ij} = \mathbf{x}_i^\top \mathbf{x}_j - \mathbb{E}\mathbf{x}_i^\top \mathbf{x}_j$ and $\mathbf{x}_i$ represents the i-th row of X. One can easily see that G_{ij} is asymptotically centred Gaussian with variance $1/p$ by CLT. Thus we can crudely estimate $G_{ij} = O_\prec(p^{-1/2})$ and $\|G\| \leq \|G\|_{\mathrm{HS}} = O_\prec(\sqrt{p}\,)$.

Then, for $\mathcal{G}_1$, we can obtain that for $z \in \mathcal{D}$,

$$\mathcal{G}_1 = (r^{-\frac{1}{2}} - z)^{-1}\Big(I_p + (r^{-\frac{1}{2}} - z)^{-1}G\Big)^{-1} = (r^{-\frac{1}{2}} - z)^{-1}I_p - (r^{-\frac{1}{2}} - z)^{-2}G + O_{\prec}\big(r^{\frac{3}{2}}p\big)$$

here with a little abuse of notation, we used $O_{\prec}\big(r^{\frac{3}{2}}p\big)$ to represent the higher order term of matrix form whose operator norm is $O_{\prec}\big(r^{\frac{3}{2}}p\big)$. Choosing ϵ sufficiently small so that $p^3n^{-\frac{1}{2}} = o(1)$. After elementary calculation, we further have that

$$\begin{aligned}&(\mathcal{G}_1)_{ij} = (r^{-\frac{1}{2}} - z)^{-1}\delta_{ij} - (r^{-\frac{1}{2}} - z)^{-2}G_{ij} + O_{\prec}(n^{-1}) = (r^{-\frac{1}{2}} - z)^{-1}\delta_{ij} + O_{\prec}(n^{-\frac{1}{2}}r^{\frac{1}{2}}),\\ &m_1(z) - (r^{-\frac{1}{2}} - z)^{-1} = O_{\prec}(r^{\frac{3}{2}}),\end{aligned} \tag{C.3}$$

which by the fact that $r^{\frac{3}{2}} \ll n^{-\frac{1}{2}}r^{\frac{1}{2}}$ indeed imply the first estimate in (4.5) for the case $l = 0$. By using the identity $z\mathcal{G}_2(z) = X^\top\mathcal{G}_1(z)X - I_p$, we also have that

$$\begin{aligned}&(z\mathcal{G}_2(z))_{i'j'} = -\delta_{i'j'} + (r^{-\frac{1}{2}} - z)^{-1}(X^\top X)_{i'j'} - (r^{-\frac{1}{2}} - z)^{-2}(X^\top GX)_{i'j'} + O_{\prec}(n^{-1}),\\ &zm_2(z) = -1 + r(1 + zm_1(z)) = -1 + r^{\frac{1}{2}}(r^{-\frac{1}{2}} - z)^{-1} + O_{\prec}(r^{\frac{3}{2}}).\end{aligned} \tag{C.4}$$

It is easy to see that $(X^\top X)_{i'j'} = (\mathbf{x}^{i'})^\top\mathbf{x}^{j'} = r^{\frac{1}{2}}\delta_{i'j'} + O_{\prec}(n^{-1/2})$, where $\mathbf{x}^{i'}$ is the i'-th column of X. Furthermore, $|(X^\top GX)_{i'j'}| = |(\mathbf{x}^{i'})^\top G\mathbf{x}^{j'}| \le \|G\|\|\mathbf{x}^{i'}\|\|\mathbf{x}^{j'}\| = O_{\prec}(n^{-1/2}p)$. We then see that

$$(z\mathcal{G}_2(z))_{i'j'} - zm_2(z)\delta_{i'j'} = O_{\prec}(n^{-\frac{1}{2}}r^{\frac{1}{2}}).$$

Thus, we can conclude the second estimate in (4.5). Next, for the two estimates in (4.6), we only need to focus on the former one in light of $X^\top\mathcal{G}_1 = \mathcal{G}_2X$ and the facts $|z| \le r^{-\frac{1}{2}}$ for $z \in \mathcal{D}$, $|z| = O(1)$ for $z \in \mathcal{D}^0$. Similarly to the above discussion, we have

$$(X^\top\mathcal{G}_1(z))_{i'i} = (r^{-\frac{1}{2}} - z)^{-1}X_{ii'} - (r^{-\frac{1}{2}} - z)^{-2}(X^\top G)_{i'i} + O_{\prec}(n^{-1}p^2r^{\frac{1}{4}}) = O_{\prec}(n^{-\frac{1}{2}}r^{\frac{1}{4}}) \tag{C.5}$$

following from the facts that $X_{ii'} = O_{\prec}(n^{-\frac{1}{2}}r^{-\frac{1}{4}})$, $|(X^\top G)_{i'i}| \le \|G\||(X^\top X)_{i'i'}|^{1/2} = O_{\prec}(r^{1/4}\sqrt{p})$, and $p^2n^{-\frac{1}{2}} = o(1)$. This proved (4.6). We then turn to the estimates in (4.7). Note that G_{ii} are i.i.d. random variables of order $O_{\prec}\big(p^{-\frac{1}{2}}\big)$, for $1 \le i \le p$. Hence by CLT, $p^{-1}\sum_{i=1}^p G_{ii}$ is crudely of order $O_{\prec}(p^{-1})$. Applying the first estimate in (C.3), we have

$$\begin{aligned}m_{1n}(z) = \frac{1}{p}\sum_{i=1}^p(\mathcal{G}_1)_{ii} &= (r^{-\frac{1}{2}} - z)^{-1} + (r^{-\frac{1}{2}} - z)^{-2}\frac{1}{p}\sum_{i=1}^p G_{ii} + O_{\prec}(n^{-1})\\ &= (r^{-\frac{1}{2}} - z)^{-1} + O_{\prec}(n^{-1}).\end{aligned}$$

The above estimate, together with the second equation in (C.3) and the estimate $r^{\frac{3}{2}} \ll n^{-1}$, yields the first estimate in (4.7). The second estimate in (4.7) can be concluded simply by using the identity $zm_{2n}(z) = -1 + r(1 + zm_{1n}(z))$ and $zm_2(z) = -1 + r(1 + zm_1(z))$, since

$$|rzm_{1n}(z) - rzm_1(z)| \prec r|z||m_{1n}(z) - m_1(z)| \prec n^{-1}r^{\frac{1}{2}}$$

uniformly for $z \in \mathcal{D}$. Particularly for $z \in \mathcal{D}^0$, since $|z| = O(1)$, the bound above can be further improved to $n^{-1}r$.

Therefore, we proved the estimates (4.5)-(4.7) uniformly for $z \in \mathcal{D}$ in the case of $l = 0$. Since $\mathcal{D}^0$ is simply a subset of $\mathcal{D}$, we trivially have the results uniformly for $z \in \mathcal{D}^0$. Now, we will proceed to the case that $l \geq 1$ by using the estimates for $z \in \mathcal{D}$.

- For the case of $l \geq 1$.

We can derive the estimates easily from the case $l = 0$ by using Cauchy integral with the radius of the contour taking value $|z - \lambda_-|/4 \asymp r^{-\frac{1}{2}}$. Note that for any $z \in \mathcal{D}^0$, the contour Γ centred at z with radius $|z - \lambda_-|/4$ still lies in the regime $\mathcal{D}$, hence all the estimates (4.5)-(4.7) hold uniformly on the contour. Moreover, we shall see that

$$\Big|\big(\mathcal{G}_1^{(l)}(z)\big)_{ij} - m_1^{(l)}(z)\delta_{ij}\Big| \asymp \Big|\oint_\Gamma \frac{\big(\mathcal{G}_1(\tilde{z})\big)_{ij} - m_1(\tilde{z})\delta_{ij}}{(\tilde{z} - z)^{l+1}}\mathrm{d}\tilde{z}\Big| \prec \frac{n^{-\frac{1}{2}}r^{\frac{1}{2}}}{|z - \lambda_-|^l} = n^{-\frac{1}{2}}r^{\frac{1+l}{2}}.$$

Similarly, we can show the error bounds for the other terms stated in (4.5)-(4.7).

Appendix D. Proofs of Lemma 3 and Proposition 2

In this section, we prove Lemma 3 and Proposition 2, which are the key technical ingredients of the proofs of our main theorem. We separate the discussion into three subsections: in the first subsection we will show the proof of Lemma 3; then followed by the proof of Proposition 2 in the second subsection; in the last subsection, we provide the proofs for some technical results in the first two subsections. In advance of the proofs, we discuss some identities regarding Stieltjes transforms $m_1(z), m_2(z)$ (see (4.2) for definitions) and list some basic identities of Green functions which will be used frequently throughout this section.

Using (4.2) and (4.3), one can easily derive the following identities

$$m_1 = -\frac{1}{z(1 + r^{-1/2}m_2)}\,, \quad 1 + zm_1 = \frac{1 + zm_2}{r}\,, \quad r^{-1/2}(zm_2)' + 1 = \frac{m_1'}{m_1^2}\,. \tag{D.1}$$

We remark that since our discussion is based on the assumption $r \equiv r_n \to r_0 \in [0, 1)$, then by definition, $\lambda_- = r^{1/2} + r^{-1/2} - 2 = O(r^{-1/2})$. This implies the support of $\nu_{\mathrm{MP},a}(\mathrm{d}x)$ for $a = 1, 2$ stays away from 0 by $O(r^{-1/2})$ distance. For the special case $z = 0$, $m_1(z)$ is well-defined and analytic at $z = 0$ since $r < 1$. More specifically, $m_1(0) = \sqrt{r}/(1 - r)$ by the first equation of (4.3). In contrast, $z = 0$ is a pole of $m_2(z)$ due to the $(1 - r)$ point mass at 0 (see MP law $\nu_{\mathrm{MP},2}(\mathrm{d}x)$ in (4.1)). However, the singularity at $z = 0$ is removable for $zm_2(z)$. We can get $zm_2(z)|_{z=0} = r - 1$ by simple calculations of the second equation of (4.2).We write $\widehat{m}_2(z) := zm_2(z)$ for simplicity. Let us simply list several results of functions in terms of $m_{1,2}$ at $z = 0$ which can checked easily from either (4.2) or (D.1).

$$m_1(0) = \frac{\sqrt{r}}{1 - r}\,, \quad m_1'(0) = \frac{r}{(1 - r)^3}\,; \tag{D.2}$$

$$\widehat{m}_2(0) = r - 1\,, \quad \widehat{m}_2'(0) = \frac{r^{3/2}}{1 - r}\,. \tag{D.3}$$

Next for the Green functions $\mathcal{G}_1$, $\mathcal{G}_2$, we have some basic and useful identities which can be easily checked by some elementary computations.

$$\mathcal{G}_1^l = \frac{1}{(l-1)!}\frac{\partial^{l-1}\mathcal{G}_1}{\partial z^{l-1}} = \frac{1}{(l-1)!}\mathcal{G}_1^{(l-1)}\,, \tag{D.4}$$

$$\mathcal{G}_1^l XX^\top = \mathcal{G}_1^{l-1} + z\mathcal{G}_1^l, \quad X^\top \mathcal{G}_1^l X = \mathcal{G}_2^l X^\top X = \mathcal{G}_2^{l-1} + z\mathcal{G}_2^l\,. \tag{D.5}$$

D.1 Proof of Lemma 3

We start with the proof of (4.12). Applying Woodbury matrix identity, from (2.4), we see that

$$\widehat{\Sigma}^{-1} = \frac{n-2}{n\sqrt{r}}\Sigma^{-\frac{1}{2}}\Big(\mathcal{G}_1(0) + \mathcal{G}_1(0)XE\mathcal{I}_2^{-1}E^\top X^\top \mathcal{G}_1(0)\Big)\Sigma^{-\frac{1}{2}}\,, \tag{D.6}$$

where we introduced the notation

$$\mathcal{I}_2 := I_2 - E^\top X^\top \mathcal{G}_1(0)XE\,.$$

Recall the definition $E = (\mathbf{e}_0, \mathbf{e}_1)$. By the second identity in (D.5) and the second estimate of (4.9), we have the estimate

$$\mathbf{u}^\top X^\top \mathcal{G}_1^a(z)X\mathbf{v} = (1+zm_2(z))^{(a-1)}\mathbf{u}^\top\mathbf{v} + O_\prec(n^{-\frac{1}{2}}r^{\frac{a}{2}}) \tag{D.7}$$

for arbitrary unit vectors $\mathbf{u}, \mathbf{v}$ and any integer $a \geq 1$. Further by $\widehat{m}_2(0) = zm_2(z)\Big|_{z=0} = r-1$, we obtain

$$\mathbf{e}_0^\top X^\top \mathcal{G}_1(0)X\mathbf{e}_1 = O_\prec(n^{-1/2}r^{1/2}), \quad 1 - \mathbf{e}_i^\top X^\top \mathcal{G}_1(0)X\mathbf{e}_i = 1 - r + O_\prec(n^{-1/2}r^{1/2})\,,\ i = 0,1\,.$$

Then,

$$\mathcal{I}_2^{-1} = \frac{1}{1-r}\,I_2 + \Delta\,, \tag{D.8}$$

where Δ represents a 2×2 matrix with $\|\Delta\| = O_\prec(n^{-1/2}r^{1/2})$. Plugging (D.8) into (D.6), we can write

$$\widehat{\Sigma}^{-1} = \frac{n-2}{n\sqrt{r}}\Sigma^{-\frac{1}{2}}\mathcal{G}_1(0)\Sigma^{-\frac{1}{2}} + \frac{n-2}{n(1-r)\sqrt{r}}\sum_{i=1,2}\Sigma^{-\frac{1}{2}}\mathcal{G}_1(0)X\mathbf{e}_i\mathbf{e}_i^\top X^\top \mathcal{G}_1(0)\Sigma^{-\frac{1}{2}} + \widehat{\Delta}\,, \tag{D.9}$$

where

$$\widehat{\Delta} = \frac{n-2}{n\sqrt{r}}\Sigma^{-\frac{1}{2}}\mathcal{G}_1(0)XE\Delta E^\top X^\top \mathcal{G}_1(0)\Sigma^{-\frac{1}{2}}\,,$$

and it is easy to check $\|\widehat{\Delta}\| \prec n^{-1/2}r^{1/2}$.

With the above preparation, we now compute the leading term of $\widehat{A}^\top\Sigma\widehat{A}$. Recall $\widehat{A} = \widehat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_d$. We have

$$\begin{aligned}\widehat{A}^\top\Sigma\widehat{A} = \hat{\boldsymbol{\mu}}_d^\top\widehat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_d &= \sqrt{r}\mathbf{v}_1^\top X^\top\Sigma^{\frac{1}{2}}\widehat{\Sigma}^{-1}\Sigma^{\frac{1}{2}}X\mathbf{v}_1 + \boldsymbol{\mu}_d^\top\widehat{\Sigma}^{-1}\boldsymbol{\mu}_d + 2r^{\frac{1}{4}}\mathbf{v}_1^\top X^\top\Sigma^{\frac{1}{2}}\widehat{\Sigma}^{-1}\boldsymbol{\mu}_d\\ &=: T_1 + T_2 + T_3\,.\end{aligned} \tag{D.10}$$

For T_1, with (D.9), we have

$$\begin{aligned} T_1 &= \frac{n-2}{n}\,\mathbf{v}_1^\top X^\top \mathcal{G}_1(0) X \mathbf{v}_1 + \frac{n-2}{n(1-r)}\,\mathbf{v}_1^\top X^\top \mathcal{G}_1(0) X \Big(\sum_{i=0,1} \mathbf{e}_i \mathbf{e}_i^\top\Big) X^\top \mathcal{G}_1(0) X \mathbf{v}_1 \\ &\qquad + \sqrt{r}\mathbf{v}_1^\top X^\top \Sigma^{\frac12} \widehat{\Delta} \Sigma^{\frac12} X \mathbf{v}_1 \\ &= r\|\mathbf{v}_1\|^2 + \frac{r^2}{1-r}\Big(\big(\mathbf{v}_1^\top \mathbf{e}_0\big)^2 + \big(\mathbf{v}_1^\top \mathbf{e}_1\big)^2\Big) + O_\prec(n^{-\frac12} r^{\frac12})\,. \qquad \text{(D.11)} \end{aligned}$$

Here in the last step, we repeatedly used the estimate (D.7) and $1 + zm_2(z)|_{z=0} = r$. In addition, for the last term of the second line of (D.11), we trivially bound it by

$$\sqrt{r}\mathbf{v}_1^\top X^\top \Sigma^{\frac12} \widehat{\Delta} \Sigma^{\frac12} X \mathbf{v}_1 \le \|\widehat{\Delta}\| \|\Sigma\| (\sqrt{r}\mathbf{v}_1^\top X^\top X \mathbf{v}_1) = O_\prec(n^{-\frac12} r^{\frac12})\,.$$

Similarly, for T_2, we have

$$\begin{aligned} T_2 &= \frac{n-2}{n\sqrt{r}}\,\mathbf{u}_1^\top\ \mathcal{G}_1(0)\mathbf{u}_1 + \frac{n-2}{n(1-r)\sqrt{r}}\,\mathbf{u}_1^\top \mathcal{G}_1(0) X \Big(\sum_{i=0,1} \mathbf{e}_i \mathbf{e}_i^\top\Big) X^\top \mathcal{G}_1(0)\mathbf{u}_1 + \mathbf{u}_1^\top \Sigma^{\frac12} \widehat{\Delta} \Sigma^{\frac12} \mathbf{u}_1 \\ &= \frac{\|\mathbf{u}_1\|^2}{1-r} + O_\prec(n^{-\frac12}\|\mathbf{u}_1\|^2)\,, \qquad \text{(D.12)} \end{aligned}$$

where we employed the shorthand notation

$$\mathbf{u}_1 := \Sigma^{-\frac12} \boldsymbol{\mu}_d\,. \qquad \text{(D.13)}$$

Here in (D.12), we applied the estimates

$$\begin{aligned} \mathbf{u}_1^\top\ \mathcal{G}_1(0)\mathbf{u}_1 &= m_1(0)\|\mathbf{u}_1\|^2 + O_\prec(n^{-\frac12} r^{\frac12} \|\mathbf{u}_1\|^2) \qquad \text{(D.14)} \\ \mathbf{u}_1^\top \mathcal{G}_1(z) X \mathbf{e}_i &= O_\prec(n^{-\frac12} r^{\frac14}\|\mathbf{u}_1\|), \quad i = 0, 1\,. \qquad \text{(D.15)} \end{aligned}$$

with the fact $m_1(0) = \sqrt{r}/(1-r)$. Next, we turn to estimate T_3. Similarly, we have

$$\begin{aligned} T_3 &= \frac{2(n-2)}{nr^{\frac14}} \mathbf{v}_1^\top X^\top \mathcal{G}_1(0)\mathbf{u}_1 + \frac{2(n-2)}{n(1-r)r^{\frac14}} \mathbf{v}_1^\top X^\top \mathcal{G}_1(0) X \Big(\sum_{i=0,1} \mathbf{e}_i \mathbf{e}_i^\top\Big) X^\top \mathcal{G}_1(0)\mathbf{u}_1 + O_\prec(n^{-\frac12} r^{\frac12}\|\mathbf{u}_1\|) \\ &= O_\prec(n^{-\frac12}\|\mathbf{u}_1\|)\,. \end{aligned}$$

Therefore, we arrive at

$$\widehat{A}^\top \widehat{\Sigma} \widehat{A} = \frac{r}{1-r}\|\mathbf{v}_1\|^2 + \frac{1}{1-r}\|\mathbf{u}_1\|^2 + O_\prec\big(n^{-\frac12}(r^{\frac12} + \|\mathbf{u}_1\|^2 + \|\mathbf{u}_1\|)\big)\,.$$

This proved (4.12).

To proceed, we estimate $\widehat{A}^\top \Sigma \widehat{A}$. By definition,

$$\widehat{A}^\top \Sigma \widehat{A} = \Big(\frac{n-2}{n\sqrt{r}}\Big)^2 \widehat{\boldsymbol{\mu}}_d^\top \Sigma^{-\frac12} H_E^{-2} \Sigma^{-\frac12} \widehat{\boldsymbol{\mu}}_d\,,$$

where we introduced the notation

$$H_E := X(I_n - EE^\top)X^\top .$$

Applying Woodbury matrix identity again, we have

$$\begin{aligned} H_E^{-2} &= \mathcal{G}_1^2(0) + \mathcal{G}_1^2(0)XE\mathcal{I}_2^{-1}E^\top X^\top \mathcal{G}_1(0) \\ &\quad + \mathcal{G}_1(0)XE\mathcal{I}_2^{-1}E^\top X^\top \mathcal{G}_1^2(0) + \Big(\mathcal{G}_1(0)XE\mathcal{I}_2^{-1}E^\top X^\top \mathcal{G}_1(0)\Big)^2 . \end{aligned} \tag{D.16}$$

Analogously to the way we deal with $\widehat{A}^\top \widehat{\Sigma}\widehat{A}$, applying the representation of $\hat{\boldsymbol{\mu}}_d$ in (2.6) and also the notation in (D.13), we can write

$$\begin{aligned} \widehat{A}^\top \Sigma \widehat{A} &= \Big(\frac{n-2}{n}\Big)^2 r^{-\frac{1}{2}} \mathbf{v}_1^\top X^\top H_E^{-2} X\mathbf{v}_1 + \Big(\frac{n-2}{n}\Big)^2 r^{-1}\mathbf{u}_1^\top H_E^{-2}\mathbf{u}_1 \\ &\quad + 2\Big(\frac{n-2}{n}\Big)^2 r^{-\frac{3}{4}} \mathbf{v}_1^\top X^\top H_E^{-2}\mathbf{u}_1 =: \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3 , \end{aligned} \tag{D.17}$$

and we analyse the RHS of the above equation term by term. First, for $\mathcal{T}_1$, substituting (D.16) and (D.8), we have

$$\begin{aligned} \mathcal{T}_1 &= \Big(\frac{n-2}{n}\Big)^2 r^{-\frac{1}{2}} \Big(\mathbf{v}_1^\top X^\top \mathcal{G}_1^2(0) X\mathbf{v}_1 + \frac{2}{1-r}\sum_{i=0,1} \big(\mathbf{v}_1^\top X^\top \mathcal{G}_1^2(0) X\mathbf{e}_i\big)\big(\mathbf{e}_i^\top X^\top \mathcal{G}_1(0)X\mathbf{v}_1\big) \\ &\quad + \frac{1}{(1-r)^2}\sum_{i,j=0,1} \big(\mathbf{v}_1^\top X^\top \mathcal{G}_1(0)X\mathbf{e}_i\big)\big(\mathbf{e}_i^\top X^\top \mathcal{G}_1^2(0)X\mathbf{e}_j\big)\big(\mathbf{e}_j^\top X^\top \mathcal{G}_1(0)X\mathbf{v}_1\big)\Big) + O_\prec(n^{-\frac{1}{2}}) \\ &= \Bigg[r^{-\frac{1}{2}}(zm_2(z))'\|\mathbf{v}_1\|^2 + \frac{2r^{-\frac{1}{2}}}{1-r}(zm_2(z))'(1+zm_2(z))\Big(\sum_{i=0,1}(\mathbf{v}_1^\top \mathbf{e}_i)^2\Big) \\ &\quad + \frac{r^{-\frac{1}{2}}}{(1-r)^2}(zm_2(z))'\Big((1+zm_2(z))\Big)^2\Big(\sum_{i=0,1}(\mathbf{v}_1^\top \mathbf{e}_i)^2\Big)\Bigg]\Bigg|_{z=0} + O_\prec(n^{-\frac{1}{2}}r^{\frac{1}{2}}) \\ &= \frac{r}{(1-r)^3}\|\mathbf{v}_1\|^2 + O_\prec(n^{-\frac{1}{2}}r^{\frac{1}{2}}) . \end{aligned} \tag{D.18}$$

Here we used the estimate (D.7) and the facts that $(zm_2(z))'\big|_{z=0} = r^{3/2}/(1-r)$, $(1+zm_2(z))\big|_{z=0} = r$ and $\sum_{i=0,1}(\mathbf{v}_1^\top \mathbf{e}_i)^2 = \|\mathbf{v}_1\|^2$ according to the definition of $\mathbf{v}_1$ in (2.6).

Next, similarly to $\mathcal{T}_1$, for $\mathcal{T}_2$, we have the estimates

$$\begin{aligned} \mathcal{T}_2 &= \Big(\frac{n-2}{n}\Big)^2 r^{-1}\Big(\mathbf{u}_1^\top \mathcal{G}_1^2(0)\mathbf{u}_1 + \frac{2}{1-r}\sum_{i=0,1}\big(\mathbf{u}_1^\top \mathcal{G}_1^2(0)X\mathbf{e}_i\big)\big(\mathbf{e}_i^\top X^\top \mathcal{G}_1(0)\mathbf{u}_1\big) \\ &\quad + \frac{1}{(1-r)^2}\sum_{i,j=0,1}\big(\mathbf{u}_1^\top \mathcal{G}_1(0)X\mathbf{e}_i\big)\big(\mathbf{e}_i^\top X^\top \mathcal{G}_1^2(0)X\mathbf{e}_j\big)\big(\mathbf{e}_j^\top X^\top \mathcal{G}_1(0)\mathbf{u}_1\big)\Big) + O_\prec(n^{-\frac{1}{2}}\|\mathbf{u}_1\|^2) \\ &= \frac{1}{(1-r)^3}\|\mathbf{u}_1\|^2 + O_\prec(n^{-\frac{1}{2}}\|\mathbf{u}_1\|^2) . \end{aligned}$$

In the last step, we applied (4.9), the second estimate of (4.10) and (D.7). Further for $\mathcal{T}_3$, we have the following estimate

$$\begin{aligned}\mathcal{T}_3 &= 2\Big(\frac{n-2}{n}\Big)^2 r^{-\frac{3}{4}}\Big(\mathbf{v}_1^\top X^\top \mathcal{G}_1^2(0)\mathbf{u}_1 + \frac{1}{1-r}\sum_{i=0,1}\big(\mathbf{v}_1^\top X^\top \mathcal{G}_1^2(0)X\mathbf{e}_i\big)\big(\mathbf{e}_i^\top X^\top \mathcal{G}_1(0)\mathbf{u}_1\big)\\ &\quad + \frac{1}{1-r}\sum_{i=0,1}\big(\mathbf{v}_1^\top X^\top \mathcal{G}_1(0)X\mathbf{e}_i\big)\big(\mathbf{e}_i^\top X^\top \mathcal{G}_1^2(0)\mathbf{u}_1\big)\\ &\quad + \frac{1}{(1-r)^2}\sum_{i,j=0}^{1}\big(\mathbf{v}_1^\top X^\top \mathcal{G}_1(0)X\mathbf{e}_i\big)\big(\mathbf{e}_i^\top X^\top \mathcal{G}_1^2(0)X\mathbf{e}_j\big)\big(\mathbf{e}_j^\top X^\top \mathcal{G}_1(0)\mathbf{u}_1\big)\Big) + O_\prec(n^{-\frac{1}{2}}\|\mathbf{u}_1\|)\\ &= O_\prec(n^{-\frac{1}{2}}\|\mathbf{u}_1\|)\,.\end{aligned}$$

Here all the summands above contain quadratic forms of $(X\mathcal{G}_1^a)$, and by (4.10), we see such quadratic forms are of order $O_\prec(n^{-\frac{1}{2}}r^{1/4+(a-1)/2}\|\mathbf{u}_1\|)$. Further with the estimate (D.7) and identities (D.3), we shall get the estimate $O_\prec(n^{-\frac{1}{2}}\|\mathbf{u}_1\|)$ for $\mathcal{T}_3$. According to the above estimates of $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$, we now see that

$$\widehat{A}^\top\Sigma\widehat{A} = \frac{r}{(1-r)^3}\|\mathbf{v}_1\|^2 + \frac{1}{(1-r)^3}\|\mathbf{u}_1\|^2 + O_\prec\big(n^{-\frac{1}{2}}(r^{\frac{1}{2}} + \|\mathbf{u}_1\|^2 + \|\mathbf{u}_1\|)\big)\,.$$

Thus we completed the proof of (4.11) by the fact that $\|\mathbf{u}_1\|^2 = \Delta_d$.

Next, we turn to prove the estimates (4.13) and (4.14). Recall the representations of $\hat{\boldsymbol{\mu}}^0$ and $\hat{\boldsymbol{\mu}}_d$ in (2.5) and (2.6), and also the notation in (D.13). Applying Woodbury matrix identity to H_E^{-1}, we can write

$$\begin{aligned}\widehat{A}^\top\boldsymbol{\mu}_d &= \frac{n-2}{n\sqrt{r}}\big(r^{\frac{1}{4}}\mathbf{v}_1^\top X^\top + \mathbf{u}_1^\top\big)H_E^{-1}\mathbf{u}_1\\ &= \frac{n-2}{n}r^{-\frac{1}{4}}\Big(\mathbf{v}_1^\top X^\top\mathcal{G}_1(0)\mathbf{u}_1 + \mathbf{v}_1^\top X^\top\mathcal{G}_1(0)XE\mathcal{I}_2^{-1}E^\top X^\top\mathcal{G}_1(0)\mathbf{u}_1\Big)\\ &\quad + \frac{n-2}{n\sqrt{r}}\Big(\mathbf{u}_1^\top\mathcal{G}_1(0)\mathbf{u}_1 + \mathbf{u}_1^\top\mathcal{G}_1(0)XE\mathcal{I}_2^{-1}E^\top X^\top\mathcal{G}_1(0)\mathbf{u}_1\Big)\,,\end{aligned}$$

and

$$\begin{aligned}\widehat{A}^\top\hat{\boldsymbol{\mu}}^0 - \widehat{A}^\top\boldsymbol{\mu}^0 = \hat{\boldsymbol{\mu}}_d^\top\widehat{\Sigma}^{-1}(\hat{\boldsymbol{\mu}}^0 - \boldsymbol{\mu}^0) &= \frac{n-2}{n\sqrt{r}}\big(r^{\frac{1}{4}}\mathbf{v}_1^\top X^\top + \mathbf{u}_1^\top\big)H_E^{-1}\Big(\sqrt{\frac{n}{n_0}}r^{\frac{1}{4}}X\mathbf{e}_0\Big)\\ &= \frac{n-2}{\sqrt{nn_0}}\Big(\mathbf{v}_1^\top X^\top\mathcal{G}_1(0)X\mathbf{e}_0 + \mathbf{v}_1^\top X^\top\mathcal{G}_1(0)XE\mathcal{I}_2^{-1}E^\top X^\top\mathcal{G}_1(0)X\mathbf{e}_0\Big)\\ &\quad + \frac{n-2}{\sqrt{nn_0}}r^{-\frac{1}{4}}\Big(\mathbf{u}_1^\top\mathcal{G}_1(0)X\mathbf{e}_0 + \mathbf{u}_1^\top\mathcal{G}_1(0)XE\mathcal{I}_2^{-1}E^\top X^\top\mathcal{G}_1(0)X\mathbf{e}_0\Big)\,.\end{aligned}$$

Similarly to the derivation of the leading term of $\widehat{A}^\top\widehat{\Sigma}^{-1}\widehat{A}$, by (4.9), (4.10) and (D.7), after elementary calculation, we arrive at

$$\widehat{A}^\top\boldsymbol{\mu}_d = \frac{1}{1-r}\boldsymbol{\mu}_d^\top\Sigma^{-1}\boldsymbol{\mu}_d + O_\prec\big(n^{-\frac{1}{2}}(\|\mathbf{u}_1\|^2 + \|\mathbf{u}_1\|)\big)$$

and

$$\widehat{A}^\top \hat{\boldsymbol{\mu}}^0 - \widehat{A}^\top \boldsymbol{\mu}^0 = \sqrt{\frac{n}{n_0}} \frac{r}{1-r} \mathbf{v}_1^\top \mathbf{e}_0 + O_\prec\big(n_0^{-\frac{1}{2}}(r^{\frac{1}{2}} + \|\mathbf{u}_1\|)\big) .$$

Finally, analogously to $\widehat{A}^\top \hat{\boldsymbol{\mu}}^0 - \widehat{A}^\top \boldsymbol{\mu}^0$, the estimates with the triple $(\boldsymbol{\mu}^0, \hat{\boldsymbol{\mu}}_0, \sqrt{n/n_0}\, \mathbf{e}_0)$ replaced by $(\boldsymbol{\mu}^1, \hat{\boldsymbol{\mu}}^1, \sqrt{n/n_1}\, \mathbf{e}_1)$ or $(\boldsymbol{\mu}_d, \hat{\boldsymbol{\mu}}_d, \mathbf{v}_1)$ can be derived similarly. Hence we skip the details and conclude the proof of Lemma 3.

D.2 Proof of Proposition 2

In this part, we show the proof of Proposition 2. First, we introduce the Green function representation of $\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0) - F(\Sigma, \boldsymbol{\mu}^0)$ based on Lemma 3 and Remark 3.

Lemma D.1 *Let $\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0)$ and $F(\Sigma, \boldsymbol{\mu}^0)$ be defined in (3.3) and (3.2), respectively. Suppose that Assumption 1 holds. Then,*

$$\begin{aligned}\widehat{F}(\widehat{\Sigma}, \hat{\boldsymbol{\mu}}^0) - F(\Sigma, \boldsymbol{\mu}^0) = \Bigg[& \frac{1-r}{2\sqrt{\hat{\boldsymbol{\mu}}_d^\top \widehat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}_d}} \bigg(\frac{1-2r}{(1-r)^4} \mathbf{v}_1^\top \big(z\mathcal{G}_2 - zm_2\big)\mathbf{v}_1 - \frac{r^{-\frac{1}{2}}}{(1-r)^2} \mathbf{v}_1^\top \Big((z\mathcal{G}_2)' - (zm_2)' \Big) \mathbf{v}_1 \\ & + \frac{r^{-\frac{1}{2}}}{(1-r)^2} \mathbf{u}_1^\top (\mathcal{G}_1 - m_1)\mathbf{u}_1 - r^{-1}\mathbf{u}_1^\top (\mathcal{G}_1^2 - m_1')\mathbf{u}_1 \\ & + \frac{2r^{-\frac{1}{4}}}{(1-r)^2} \mathbf{u}_1^\top \mathcal{G}_1 X \mathbf{v}_1 - \frac{2r^{-\frac{3}{4}}}{1-r} \mathbf{u}_1^\top \mathcal{G}_1^2 X \mathbf{v}_1 \bigg) \Phi^{-1}(1-\alpha) \\ & + \sqrt{\frac{n}{n_0}} \Big(\frac{1}{(1-r)^2} \mathbf{v}_1^\top (z\mathcal{G}_2 - zm_2)\mathbf{e}_0 + \frac{r^{-\frac{1}{4}}}{1-r} \mathbf{u}_1^\top \mathcal{G}_1 X \mathbf{e}_0 \Big) \Bigg] \Bigg|_{z=0} \\ & + O_\prec\big(n^{-1}(r^{\frac{1}{2}} + \Delta_d^{\frac{1}{2}})\big) . \end{aligned} \tag{D.19}$$

Remark 5 *Here we emphasize again that $z = 0$ is a removable singularity of $z\mathcal{G}_2(z)$ and $zm_2(z)$. Additionally, $z\mathcal{G}_2(z) \neq 0$ and $zm_2(z) \neq 0$ when $z = 0$ (see (D.3)). By (4.12), (4.9) and (4.10), it is not hard to see that the factor before $\Phi^{-1}(1-\alpha)$ on the RHS of (D.19) is of order $O_\prec(n^{-1/2}\Delta_d^{1/2})$. Similarly, the term in the fourth line of (D.19) is also crudely bounded by $O_\prec(n^{-1/2}\Delta_d^{1/2})$.*

Here to the rest of this subsection, we will adopt the notation $(M)_{\mathbf{uv}}$ as the quadratic form $\mathbf{u}^* M \mathbf{v}$ for arbitrary two column vectors $\mathbf{u}, \mathbf{v}$ of dimension a, b, respectively, and any $a \times b$ matrix M. In light of Lemma D.1 and Remark 5, it suffices to study the joint distribution of the following terms with appropriate scalings which make them order one random variables,

$$\begin{aligned} & \frac{\sqrt{n}}{\sqrt{r}} \big(z\mathcal{G}_2 - zm_2\big)_{\bar{\mathbf{v}}_1\bar{\mathbf{v}}_1}, \ \frac{\sqrt{n}}{r} \Big((z\mathcal{G}_2)' - (zm_2)'\Big)_{\bar{\mathbf{v}}_1\bar{\mathbf{v}}_1}, \ \frac{\sqrt{n}}{\sqrt{r}} (\mathcal{G}_1 - m_1)_{\bar{\mathbf{u}}_1\bar{\mathbf{u}}_1}, \ \frac{\sqrt{n}}{r} (\mathcal{G}_1^2 - m_1')_{\bar{\mathbf{u}}_1\bar{\mathbf{u}}_1}, \\ & \sqrt{n} r^{-\frac{1}{4}} (\mathcal{G}_1 X)_{\bar{\mathbf{u}}_1\bar{\mathbf{v}}_1}, \ \sqrt{n} r^{-\frac{3}{4}} (\mathcal{G}_1^2 X)_{\bar{\mathbf{u}}_1\bar{\mathbf{v}}_1}, \ \frac{\sqrt{n}}{\sqrt{r}} (z\mathcal{G}_2 - zm_2)_{\bar{\mathbf{v}}_1\mathbf{e}_0}, \ \sqrt{n} r^{-\frac{1}{4}} (\mathcal{G}_1 X)_{\bar{\mathbf{u}}_1\mathbf{e}_0} . \end{aligned} \tag{D.20}$$

Here we adopt the notation $\bar{\mathbf{u}}$ to denote the normalized version of a generic vector $\mathbf{u}$, i.e.

$$\bar{\mathbf{u}} = \begin{cases} \frac{\mathbf{u}}{\|\mathbf{u}\|}, & \text{if } \|\mathbf{u}\| \neq 0\,; \\ 0, & \text{otherwise}\,. \end{cases}$$

And for a fixed deterministic column vector $\mathbf{c} := \big(c_{10}, \cdots, c_{14}, c_{20}, c_{21}, c_{22}\big)^\top \in \mathbb{R}^8$, we define for $z \in \mathcal{D}$

$$\begin{aligned} \mathcal{P} \equiv \mathcal{P}(\mathbf{c}, z) := {} & \frac{\sqrt{n}}{\sqrt{r}} c_{10}(\mathcal{G}_1 - m_1)_{\bar{\mathbf{u}}_1\bar{\mathbf{u}}_1} + \frac{\sqrt{n}}{r} c_{11}(\mathcal{G}_1^2 - m_1')_{\bar{\mathbf{u}}_1\bar{\mathbf{u}}_1} \\ & + \frac{\sqrt{n}}{r^{\frac{1}{4}}} c_{12}(\mathcal{G}_1 X)_{\bar{\mathbf{u}}_1\bar{\mathbf{v}}_1} + \frac{\sqrt{n}}{r^{\frac{1}{4}}} c_{13}(\mathcal{G}_1 X)_{\bar{\mathbf{u}}_1\mathbf{e}_0} + \frac{\sqrt{n}}{r^{\frac{3}{4}}} c_{14}(\mathcal{G}_1^2 X)_{\bar{\mathbf{u}}_1\bar{\mathbf{v}}_1} \\ & + \frac{\sqrt{n}}{\sqrt{r}} c_{20}\big(z\mathcal{G}_2 - zm_2\big)_{\bar{\mathbf{v}}_1\bar{\mathbf{v}}_1} + \frac{\sqrt{n}}{\sqrt{r}} c_{21}(z\mathcal{G}_2 - zm_2)_{\bar{\mathbf{v}}_1\mathbf{e}_0} + \frac{\sqrt{n}}{r} c_{22}\Big((z\mathcal{G}_2)' - (zm_2)'\Big)_{\bar{\mathbf{v}}_1\bar{\mathbf{v}}_1}. \end{aligned} \tag{D.21}$$

Further we define $\mathcal{M} \equiv \mathcal{M}(z)$ to be a 8-by-8 block diagonal matrix such that $\mathcal{M} = \operatorname{diag}(\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3)$, and the main-diagonal blocks $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ are all symmetric matrices with dimension $2, 3, 3$, respectively. The entrywise definition of the diagonal blocks are given below.

With certain abuse of notation, in this part, let us use $\mathcal{M}_a(i,j)$ to denote the (i,j)-th entry of matrix $\mathcal{M}_a, a = 1, 2, 3$. For the matrix $\mathcal{M}_1$, it is defined entrywise by

$$\begin{aligned} & \mathcal{M}_1(1,1) = 2r^{-\frac{3}{2}} m_1^2 (zm_1)'\,, \qquad \mathcal{M}_1(1,2) = r^{-2} m_1^2 (zm_1)'' + 2r^{-2} m_1 m_1' (zm_1)'\,, \\ & \mathcal{M}_1(2,2) = 2r^{-\frac{5}{2}} \Big(\frac{m_1^2 (zm_1)'''}{3!} + m_1 m_1' (zm_1)'' + (m_1')^2 (zm_1)' \Big)\,. \end{aligned}$$

The entries of $\mathcal{M}_2$ are given by

$$\begin{aligned} \mathcal{M}_2(1,1) &= -\frac{m_1'(zm_2)}{r(1+\sqrt{r}m_1)}\,, \qquad \mathcal{M}_2(1,2) = \frac{m_1'(zm_2)}{r(1+\sqrt{r}m_1)}\sqrt{\frac{n_1}{n}}\,, \\ \mathcal{M}_2(1,3) &= \frac{1}{2}\Big[-\frac{m_1''(zm_2)}{r^{\frac{3}{2}}(1+\sqrt{r}m_1)} - \frac{m_1'(zm_2)'}{r^{\frac{3}{2}}(1+\sqrt{r}m_1)} + \frac{(m_1')^2(zm_2)}{r(1+\sqrt{r}m_1)^2} \Big]\,, \\ \mathcal{M}_2(2,2) &= -\frac{m_1'(zm_2)}{r(1+\sqrt{r}m_1)}\,, \\ \mathcal{M}_2(2,3) &= -\frac{1}{2}\Big[-\frac{m_1''(zm_2)}{r^{\frac{3}{2}}(1+\sqrt{r}m_1)} - \frac{m_1'(zm_2)'}{r^{\frac{3}{2}}(1+\sqrt{r}m_1)} + \frac{(m_1')^2(zm_2)}{r(1+\sqrt{r}m_1)^2} \Big]\sqrt{\frac{n_1}{n}}\,, \\ \mathcal{M}_2(3,3) &= -\frac{1}{r^2(1+\sqrt{r}m_1)}\Big(\frac{m_1'''(zm_2)}{3!} + \frac{m_1''(zm_2)'}{2} \Big) + \frac{m_1'}{r^{\frac{3}{2}}(1+\sqrt{r}m_1)^2}\Big(\frac{m_1''(zm_2)}{2} + m_1'(zm_2)' \Big)\,. \end{aligned}$$

Further, we define $\mathcal{M}_3$ entrywise by

$$\begin{aligned}
\mathcal{M}_3(1,1) &= -\frac{2(zm_2)'(zm_2)}{r^{\frac{3}{2}}(1+\sqrt{r}m_1)}\,, \qquad \mathcal{M}_3(1,2) = \frac{2(zm_2)'(zm_2)}{r^{\frac{3}{2}}(1+\sqrt{r}m_1)}\sqrt{\frac{n_1}{n}}\,,\\
\mathcal{M}_3(1,3) &= -\frac{(zm_2)''(zm_2)}{r^2(1+\sqrt{r}m_1)} + \frac{m_1'(zm_2)'(zm_2)}{r^{\frac{3}{2}}(1+\sqrt{r}m_1)^2} - \frac{\big((zm_2)'\big)^2}{r^2(1+\sqrt{r}m_1)}\,,\\
\mathcal{M}_3(2,2) &= -\frac{(zm_2)'(zm_2)}{r^{\frac{3}{2}}(1+\sqrt{r}m_1)}\Big(1+\frac{n_1}{n}\Big)\,,\\
\mathcal{M}_3(2,3) &= \Big(-\frac{(zm_2)''(zm_2)}{r^2(1+rm_1)} + \frac{m_1'(zm_2)'(zm_2)}{r^{\frac{3}{2}}(1+rm_1)^2} - \frac{\big((zm_2)'\big)^2}{r^2(1+rm_1)}\Big)\Big(-\sqrt{\frac{n_1}{n}}\Big)\,,\\
\mathcal{M}_3(3,3) &= 2\Big[-\frac{1}{r^{\frac{5}{2}}(1+\sqrt{r}m_1)}\Big(\frac{(zm_2)'''(zm_2)}{3!} + \frac{(zm_2)''(zm_2)'}{2}\Big)\\
&\qquad + \frac{m_1'}{r^2(1+\sqrt{r}m_1)^2}\Big(\frac{(zm_2)''(zm_2)}{2} + \big((zm_2)'\big)^2\Big)\Big]\,.
\end{aligned}$$

Next, we set

$$z := \mathrm{i}n^{-K} \tag{D.22}$$

for some sufficiently large constant $K > 0$. This setting allows us to use the high probability bounds for the quadratic forms of $\mathcal{G}_1^a, (z\mathcal{G}_2)^{(a)}, (X^\top \mathcal{G}_1^a)$ for $a = 0, 1$, even when we estimate their moments. To see this, first we can always bound those quadratic forms deterministically by $(\Im z)^{-s}$ for some fixed $s > 0$, up to some constant. Then according to Lemma 2 (ii) and Proposition 1 with Remark 2, we get that the high probability bound in Remark 2 can be directly applied in calculations of the expectations.

With all the above notations, we introduce the following proposition.

Proposition D.1 *Let $\mathcal{P}$ be defined above and z given in (D.22). Denote by $\varphi_n(\cdot)$ the characteristic function of $\mathcal{P}$. Suppose that $p/n \to [0,1)$. Then, for $|t| \ll n^{1/2}$,*

$$\varphi_n'(t) = -\big(\mathbf{c}^\top \mathcal{M}\mathbf{c}\big)t\varphi_n(t) + O_\prec((|t|+1)n^{-\frac{1}{2}})\,.$$

The proof of Proposition D.1 will be postponed. With the aid of Lemma D.1 and Proposition D.1, we can now finish the proof of Proposition 2.

Proof (Proof of Proposition 2) First by Proposition D.1, we claim that the random vector

$$\begin{aligned}
\Bigg(&\frac{\sqrt{n}}{\sqrt{r}}(\mathcal{G}_1 - m_1)_{\bar{\mathbf{u}}_1\bar{\mathbf{u}}_1},\ \frac{\sqrt{n}}{r}(\mathcal{G}_1^2 - m_1')_{\bar{\mathbf{u}}_1\bar{\mathbf{u}}_1},\ \frac{\sqrt{n}}{r^{\frac{1}{4}}}(\mathcal{G}_1 X)_{\bar{\mathbf{u}}_1\bar{\mathbf{v}}_1},\ \frac{\sqrt{n}}{r^{\frac{1}{4}}}(\mathcal{G}_1 X)_{\bar{\mathbf{u}}_1\mathbf{e}_0},\ \frac{\sqrt{n}}{r^{\frac{3}{4}}}(\mathcal{G}_1^2 X)_{\bar{\mathbf{u}}_1\bar{\mathbf{v}}_1},\\
&\frac{\sqrt{n}}{\sqrt{r}}\big(z\mathcal{G}_2 - zm_2\big)_{\bar{\mathbf{v}}_1\bar{\mathbf{v}}_1},\ \frac{\sqrt{n}}{\sqrt{r}}(z\mathcal{G}_2 - zm_2)_{\bar{\mathbf{v}}_1\mathbf{e}_0},\ \frac{\sqrt{n}}{r}\Big((z\mathcal{G}_2)' - (zm_2)'\Big)_{\bar{\mathbf{v}}_1\bar{\mathbf{v}}_1}\Bigg)
\end{aligned} \tag{D.23}$$

is asymptotically Gaussian with mean $\mathbf{0}$ and covariance matrix $\mathcal{M}$ at $z = 0$. To see this, we only need to claim that $\mathcal{P}$ is asymptotically normal with mean 0 and variance $\mathbf{c}^\top\mathcal{M}\mathbf{c}$ due to the arbitrariness of the fixed vector $\mathbf{c}$. Let us denote by $\varphi_0(t)$ the characteristic function of standard normal distribution with mean 0 and variance $\mathbf{c}^\top\mathcal{M}\mathbf{c}$ which takes the expression $\varphi_0(t) = \exp\{-\big(\mathbf{c}^\top\mathcal{M}\mathbf{c}\big)t^2/2\}$. According to Proposition D.1, for $|t| \ll n^{1/2}$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\varphi_n(t)}{\varphi_0(t)} = \frac{\varphi_n'(t) + \big(\mathbf{c}^\top\mathcal{M}\mathbf{c}\big)t\varphi_n(t)}{\varphi_0(t)} = O_{\prec}\Big((|t|+1)e^{\big(\mathbf{c}^\top\mathcal{M}\mathbf{c}\big)t^2/2}n^{-\frac{1}{2}}\Big)\,.$$

Notice the fact $\varphi(0)/\varphi_0(0) = 1$, we shall have

$$\frac{\varphi_n(t)}{\varphi_0(t)} - 1 = \begin{cases} O_{\prec}\Big(e^{\big(\mathbf{c}^\top\mathcal{M}\mathbf{c}\big)t^2/2}n^{-\frac{1}{2}}\Big), & 1 < |t| \ll \sqrt{n}\,; \\ O_{\prec}(|t|n^{-\frac{1}{2}}), & |t| \le 1\,. \end{cases}$$

This further implies that

$$\varphi_n(t) = \varphi_0(t) + O_{\prec}(n^{-\frac{1}{2}}),\ \text{for } 1 < |t| \ll \sqrt{n}; \qquad \varphi_n(t) = \varphi_0(t) + O_{\prec}(|t|n^{-\frac{1}{2}}),\ \text{for } |t| \le 1\,. \tag{D.24}$$

We can then conclude the asymptotical distribution of $\mathcal{P}$.

Recall the Green function representation in (D.19). Set

$$\begin{aligned} \mathit{\Theta}_\alpha := \Bigg[&\frac{(1-r)\sqrt{n}}{2\sqrt{\hat{\boldsymbol{\mu}}_d^\top\widehat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_d}}\bigg(\frac{1-2r}{(1-r)^4}\big(z\mathcal{G}_2 - zm_2\big)_{\mathbf{v}_1\mathbf{v}_1} - \frac{1}{r^{\frac{1}{2}}(1-r)^2}\Big((z\mathcal{G}_2)' - (zm_2)'\Big)_{\mathbf{v}_1\mathbf{v}_1} \\ &\qquad + \frac{1}{r^{\frac{1}{2}}(1-r)^2}(\mathcal{G}_1 - m_1)_{\mathbf{u}_1\mathbf{u}_1} - \frac{1}{r}(\mathcal{G}_1^2 - m_1')_{\mathbf{u}_1\mathbf{u}_1} \\ &\qquad + \frac{2}{r^{\frac{1}{4}}(1-r)^2}(\mathcal{G}_1X)_{\mathbf{u}_1\mathbf{v}_1} - \frac{2}{r^{\frac{3}{4}}(1-r)}(\mathcal{G}_1^2X)_{\mathbf{u}_1\mathbf{v}_1}\bigg)\,\Phi^{-1}(1-\alpha) \\ &+ \frac{n}{\sqrt{n_0}}\Big(\frac{1}{(1-r)^2}(z\mathcal{G}_2 - zm_2)_{\mathbf{v}_1\mathbf{e}_0} + \frac{1}{r^{\frac{1}{4}}(1-r)}(\mathcal{G}_1X)_{\mathbf{u}_1\mathbf{e}_0}\Big)\Bigg]\Big/\sqrt{\Big((1-r)\hat{\boldsymbol{\mu}}_d^\top\widehat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_d - \frac{n^2r}{n_0n_1}\Big)}\,, \end{aligned}$$

which is a linear combination of the components of the vector in (D.23). Therefore by elementary calculations of the quadratic form of $\mathcal{M}$ with the identities

$$\begin{aligned} &m_1(0) = \frac{\sqrt{r}}{1-r}, \quad m_1'(0) = \frac{r}{(1-r)^3}, \quad m_1''(0) = \frac{2r^{\frac{3}{2}}(1+r)}{(1-r)^5}, \quad m_1'''(0) = \frac{6r^2(1+3r+r^2)}{(1-r)^7} \\ &\widehat{m}_2(0) := \big(zm_2(z)\big)\Big|_{z=0} = r-1, \quad \widehat{m}_2'(0) = \frac{r^{\frac{3}{2}}}{1-r}, \quad \widehat{m}_2''(0) = \frac{2r^2}{(1-r)^3}, \quad \widehat{m}_2'''(0) = \frac{6r^{\frac{5}{2}}(1+r)}{(1-r)^5}, \end{aligned}$$

together with the estimate

$$\frac{\boldsymbol{\mu}_d^\top\Sigma^{-1}\boldsymbol{\mu}_d}{(1-r)\hat{\boldsymbol{\mu}}_d^\top\widehat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_d - \frac{n^2r}{n_0n_1}} = 1 + O_{\prec}(n^{-\frac{1}{2}})$$

which follows from Lemma 3, we can finally prove (4.17) and the fact $\mathit{\Theta}_\alpha \simeq \mathcal{N}(0,\widehat{V})$.

In the end, we show the convergence rate of $\varTheta_\alpha$ again using Proposition D.1. It suffices to obtain the convergence rate of the general form of linear combination, i.e. $\mathcal{P}$. We follow the derivations for Berry-Esseen bound, more precisely, by Esseen's inequality, we have

$$\sup_{x\in\mathbf{R}}\big|F_n(x)-F_0(x)\big| \le C_1\int_0^T \frac{|\varphi_n(t)-\varphi_0(t)|}{t}\mathrm{d}t+\frac{C_2}{T}$$

for some fixed constants $C_1, C_2 > 0$. Here we use $F_n(x), F_0(x)$ to denote the distribution functions of $\mathcal{P}$ and centred normal distribution with variance $\mathbf{c}^\top\mathcal{M}\mathbf{c}$, respectively. Applying (D.24), and choose $T=\sqrt{n}$, we then get

$$\sup_{x\in\mathbf{R}}\big|F_n(x)-F_0(x)\big| \le C_1\int_1^T t^{-1}O_\prec(n^{-\frac12})\mathrm{d}t + C_1\int_0^1 t^{-1}O_\prec(|t|n^{-\frac12})\mathrm{d}t + C_2 n^{-\frac12} = O_\prec(n^{-\frac12})\,.$$

This indicates that the convergence rate of $\mathcal{P}$ is $O_\prec(n^{-\frac12})$, and hence the same rate applies to $\varTheta_\alpha$. ■

Remark 6 *The arguments of the convergence rate of $\widetilde{\varTheta}_\alpha$ of Remark 4, which leads to the high probability bound in Corollary 1 is actually the same, since $\widetilde{\varTheta}_\alpha$ again takes the form of $\mathcal{P}$ with appropriate* $\mathbf{c}$.

D.3 Proofs of Lemma D.1 and Proposition D.1

In the last subsection, we prove the technical results from Section D.2, i.e., Lemma D.1 and Proposition D.1.

Proof (Proof of Lemma D.1)

Recall the definitions of $\widehat{F}(\widehat{\Sigma},\hat{\boldsymbol{\mu}}^0)$ and $F(\Sigma,\boldsymbol{\mu}^0)$ in (3.3) and (3.2). In light of Lemma 3, it suffices to further identify the differences $\frac{1}{(1-r)^2}\widehat{A}^\top\widehat{\Sigma}\widehat{A}-\widehat{A}^\top\Sigma\widehat{A}$ and $\widehat{A}^\top\hat{\boldsymbol{\mu}}^0-\sqrt{\frac{n}{n_0}}\frac{r}{1-r}\mathbf{v}_1^\top\mathbf{e}_0-\widehat{A}^\top\boldsymbol{\mu}^0$. We start with the first term. We write

$$\begin{aligned}\frac{1}{(1-r)^2}\widehat{A}^\top\widehat{\Sigma}\widehat{A}-\widehat{A}^\top\Sigma\widehat{A} &= \Big[\frac{n-2}{n(1-r)^2}\mathbf{v}_1^\top X^\top H_E^{-1}X\mathbf{v}_1-\Big(\frac{n-2}{n}\Big)^2 r^{-\frac12}\mathbf{v}_1^\top X^\top H_E^{-2}X\mathbf{v}_1\Big]\\ &\quad+\Big[\frac{n-2}{n(1-r)^2\sqrt{r}}\mathbf{u}_1^\top H_E^{-1}\mathbf{u}_1-\Big(\frac{n-2}{n\sqrt{r}}\Big)^2\mathbf{u}_1^\top H_E^{-2}\mathbf{u}_1\Big]\\ &\quad+2\Big[\frac{n-2}{n(1-r)^2 r^{\frac14}}\mathbf{v}_1^\top X^\top H_E^{-1}\mathbf{u}_1-\Big(\frac{n-2}{n}\Big)^2 r^{-\frac34}\mathbf{v}_1^\top X^\top H_E^{-2}\mathbf{u}_1\Big]\\ &=: D_1+D_2+D_3\,,\end{aligned}$$

in which we used (2.4), (D.10), (D.17), and the shorthand notation $\mathbf{u}_1=\Sigma^{-\frac12}\boldsymbol{\mu}_d$. In the sequel, we estimate D_1, D_2, D_3 term by term. Before we commence the details, we first continue with (D.8) to seek for the explicit form of one higher order term by resolvent expansion formula,

$$\mathcal{I}_2^{-1}=\frac{1}{1-r}I_2+\frac{1}{(1-r)^2}\mathbf{\Delta}+O_\prec(n^{-1}r)\,, \tag{D.25}$$

where

$$\boldsymbol{\Delta} = \Big(E^\top \big(z\mathcal{G}_2(z) - zm_2(z) I_p \big) E \Big)\Big|_{z=0} \,,$$

and $\|\boldsymbol{\Delta}\| = O_\prec(n^{-\frac{1}{2}} r^{\frac{1}{2}})$ by (4.9). Here in (D.25) $O_\prec(n^{-1}r)$ represents an error matrix which is stochastically bounded by r/n in operator norm. We remark here that the above estimate will be frequently used in the following calculations.

Let us start with D_1. Similarly to (D.6), by applying Woodbury matrix identity, we get

$$\begin{aligned} D_1 = {} & \frac{n-2}{n(1-r)^2} \Big(\mathbf{v}_1^\top X^\top \mathcal{G}_1 X \mathbf{v}_1 + \mathbf{v}_1^\top X^\top \mathcal{G}_1 X E \mathcal{I}_2^{-1} E^\top X^\top \mathcal{G}_1 X \mathbf{v}_1 \Big) \\ & - \Big(\frac{n-2}{n} \Big)^2 r^{-\frac{1}{2}} \Big(\mathbf{v}_1^\top X^\top \mathcal{G}_1^2 X \mathbf{v}_1 + 2 \mathbf{v}_1^\top X^\top \mathcal{G}_1^2 X E \mathcal{I}_2^{-1} E^\top X^\top \mathcal{G}_1 X \mathbf{v}_1 \Big) \\ & - \Big(\frac{n-2}{n} \Big)^2 r^{-\frac{1}{2}} \mathbf{v}_1^\top X^\top \mathcal{G}_1 X E \mathcal{I}_2^{-1} E^\top X^\top \mathcal{G}_1^2 X E \mathcal{I}_2^{-1} E^\top X^\top \mathcal{G}_1 X \mathbf{v}_1 \,. \end{aligned}$$

Hereafter, for brevity, we drop the z-dependence from the notations $\mathcal{G}_1(z)$, $\mathcal{G}_2(z)$ and $m_1(z), m_2(z)$ and set $z = 0$ but omit this fact from the notations. Recall (D.11) and (D.18). Analogously, we can compute

$$\begin{aligned} D_1 = {} & \frac{1}{(1-r)^2} \mathbf{v}_1^\top \big(z\mathcal{G}_2 - zm_2 \big) \mathbf{v}_1 - r^{-\frac{1}{2}} \mathbf{v}_1^\top \Big(\big(z\mathcal{G}_2 \big)' - (zm_2)' \Big) \mathbf{v}_1 + \frac{2(1+zm_2)}{(1-r)^3} \mathbf{v}_1^\top (z\mathcal{G}_2 - zm_2) E E^\top \mathbf{v}_1 \\ & + \frac{(1+zm_2)^2}{(1-r)^4} \mathbf{v}_1^\top E E^\top (z\mathcal{G}_2 - zm_2(z)) E E^\top \mathbf{v}_1 - \frac{2(1+zm_2)}{(1-r)\sqrt{r}} \mathbf{v}_1^\top \Big((z\mathcal{G}_2)' - (zm_2)' \Big) E E^\top \mathbf{v}_1 \\ & - \frac{2(zm_2)'}{(1-r)\sqrt{r}} \mathbf{v}_1^\top E E^\top (z\mathcal{G}_2 - zm_2) \mathbf{v}_1 - \frac{2(zm_2)'(1+zm_2)}{(1-r)^2\sqrt{r}} \mathbf{v}_1^\top E E^\top (z\mathcal{G}_2 - zm_2(z)) E E^\top \mathbf{v}_1 \\ & - \frac{2(1+zm_2)(zm_2)'}{(1-r)^2\sqrt{r}} \mathbf{v}_1^\top (z\mathcal{G}_2 - zm_2) E E^\top \mathbf{v}_1 - \frac{(1+zm_2)^2}{(1-r)^2\sqrt{r}} \mathbf{v}_1^\top E E^\top \Big((z\mathcal{G}_2)' - (zm_2)' \Big) E E^\top \mathbf{v}_1 \\ & - \frac{2(1+zm_2)^2(zm_2)'}{(1-r)^3\sqrt{r}} \mathbf{v}_1^\top E E^\top \big(z\mathcal{G}_2 - zm_2(z) \big) E E^\top \mathbf{v}_1 + O_\prec(n^{-1}r) \\ = {} & \frac{1-2r}{(1-r)^4} \mathbf{v}_1^\top \big(z\mathcal{G}_2 - zm_2 \big) \mathbf{v}_1 - \frac{1}{(1-r)^2\sqrt{r}} \mathbf{v}_1^\top \Big(\big(z\mathcal{G}_2 \big)' - (zm_2)' \Big) \mathbf{v}_1 + O_\prec(n^{-1}r) \,. \end{aligned} \tag{D.26}$$

Next, we turn to estimate D_2. Similarly to D_1, by Woodbury matrix identity, we have

$$\begin{aligned} D_2 = {} & \frac{n-2}{n(1-r)^2\sqrt{r}} \Big(\mathbf{u}_1^\top \mathcal{G}_1 \mathbf{u}_1 + \mathbf{u}_1^\top \mathcal{G}_1 X E \mathcal{I}_2^{-1} E^\top X^\top \mathcal{G}_1 \mathbf{u}_1 \Big) \\ & - \Big(\frac{n-2}{n\sqrt{r}} \Big)^2 \Big(\mathbf{u}_1^\top \mathcal{G}_1^2 \mathbf{u}_1 + 2 \mathbf{u}_1^\top \mathcal{G}_1^2 X E \mathcal{I}_2^{-1} E^\top X^\top \mathcal{G}_1 \mathbf{u}_1 \Big) \\ & - \Big(\frac{n-2}{n\sqrt{r}} \Big)^2 \mathbf{u}_1^\top \mathcal{G}_1 X E \mathcal{I}_2^{-1} E^\top X^\top \mathcal{G}_1^2 X E \mathcal{I}_2^{-1} E^\top X^\top \mathcal{G}_1 \mathbf{u}_1 \,. \end{aligned}$$

Then, by (D.25), it is not hard to derive that

$$
\begin{aligned}
D_2 &= \frac{r^{-\frac{1}{2}}}{(1-r)^2}\Big(\mathbf{u}_1^\top \mathcal{G}_1 \mathbf{u}_1 + \frac{1}{1-r}\sum_{i=0}^{1}(\mathbf{u}_1^\top \mathcal{G}_1 X\mathbf{e}_i)^2 + \frac{1}{(1-r)^2}\sum_{i,j=0}^{1}(\mathbf{u}_1^\top \mathcal{G}_1 X\mathbf{e}_i)(\mathbf{u}_1^\top \mathcal{G}_1 X\mathbf{e}_j)\big(\mathbf{e}_i^\top(z\mathcal{G}_2 - zm_2)\mathbf{e}_j\big)\Big)\\
&\quad - \frac{1}{r}\Big(\mathbf{u}_1^\top \mathcal{G}_1^2 \mathbf{u}_1 + \frac{2}{1-r}\sum_{i=0}^{1}(\mathbf{u}_1^\top \mathcal{G}_1 X\mathbf{e}_i)(\mathbf{u}_1^\top \mathcal{G}_1^2 X\mathbf{e}_i) + \frac{1}{(1-r)^2}\sum_{i,j=0}^{1}(\mathbf{u}_1^\top \mathcal{G}_1^2 X\mathbf{e}_i)(\mathbf{u}_1^\top \mathcal{G}_1 X\mathbf{e}_j)\big(\mathbf{e}_i^\top(z\mathcal{G}_2 - zm_2)\mathbf{e}_j\big)\Big)\\
&\quad - \frac{1}{(1-r)^2 r}\sum_{i,j=0}^{1}(\mathbf{u}_1^\top \mathcal{G}_1 X\mathbf{e}_i)(\mathbf{u}_1^\top \mathcal{G}_1 X\mathbf{e}_j)\big(\mathbf{e}_i^\top(z\mathcal{G}_2)'\mathbf{e}_j\big)\\
&\quad - \frac{2}{(1-r)^3 r}\sum_{i,j,k=0}^{1}(\mathbf{u}_1^\top \mathcal{G}_1 X\mathbf{e}_i)(\mathbf{u}_1^\top \mathcal{G}_1 X\mathbf{e}_k)\big(\mathbf{e}_i^\top(z\mathcal{G}_2)'\mathbf{e}_j\big)\big(\mathbf{e}_j^\top(z\mathcal{G}_2 - zm_2)\mathbf{e}_k\big) + O_\prec(n^{-1}\|\mathbf{u}_1\|^2)\\
&= \frac{1}{(1-r)^2\sqrt{r}}\mathbf{u}_1^\top(\mathcal{G}_1 - m_1)\mathbf{u}_1 - r^{-1}\mathbf{u}_1^\top(\mathcal{G}_1^2 - m_1')\mathbf{u}_1 + O_\prec(n^{-1}\|\mathbf{u}_1\|^2)\,, \qquad \text{(D.27)}
\end{aligned}
$$

where in the last step we applied the estimate $\mathbf{u}_1^\top \mathcal{G}_1^a X\mathbf{e}_i = O_\prec(n^{-1/2}r^{1/4+(a-1)/2}\|\mathbf{u}_1\|), a = 1,2$ and $\mathbf{e}_i^\top(z\mathcal{G}_2)'\mathbf{e}_j = O_\prec(r^{3/2})$ which follow from (4.10) and (4.9). Further, we also used the trivial identity $m_1(0)\sqrt{r}/(1-r)^2 = m_1'(0)$.

Next, we estimate D_3 as follows,

$$
\begin{aligned}
D_3 &= \frac{2(n-2)}{n(1-r)^2 r^{\frac{1}{4}}}\Big(\mathbf{v}_1^\top X^\top \mathcal{G}_1\mathbf{u}_1 + \mathbf{v}_1^\top X^\top \mathcal{G}_1 X E \mathcal{I}_2^{-1} E^\top X^\top \mathcal{G}_1 \mathbf{u}_1\Big)\\
&\quad - 2\Big(\frac{n-2}{n}\Big)^2 r^{-\frac{3}{4}}\Big(\mathbf{v}_1^\top X^\top \mathcal{G}_1^2\mathbf{u}_1 + \mathbf{v}_1^\top X^\top \mathcal{G}_1^2 X E\mathcal{I}_2^{-1}E^\top X^\top \mathcal{G}_1\mathbf{u}_1 + \mathbf{v}_1^\top X^\top \mathcal{G}_1 X E\mathcal{I}_2^{-1}E^\top X^\top \mathcal{G}_1^2\mathbf{u}_1\Big)\\
&\quad - 2\Big(\frac{n-2}{n}\Big)^2 r^{-\frac{3}{4}}\mathbf{v}_1^\top X^\top \mathcal{G}_1 X E\mathcal{I}_2^{-1}E^\top X^\top \mathcal{G}_1^2 X E\mathcal{I}_2^{-1}E^\top X^\top \mathcal{G}_1\mathbf{u}_1\\
&= \frac{2}{(1-r)^2 r^{\frac{1}{4}}}\Big(\mathbf{v}_1^\top X^\top \mathcal{G}_1\mathbf{u}_1 + \frac{1}{1-r}\mathbf{v}_1^\top X^\top \mathcal{G}_1 X E E^\top X^\top \mathcal{G}_1\mathbf{u}_1\Big)\\
&\quad - 2r^{-\frac{3}{4}}\Big(\mathbf{v}_1^\top X^\top \mathcal{G}_1^2\mathbf{u}_1 + \frac{1}{1-r}\mathbf{v}_1^\top X^\top \mathcal{G}_1^2 X E E^\top X^\top \mathcal{G}_1\mathbf{u}_1 + \frac{1}{1-r}\mathbf{v}_1^\top X^\top \mathcal{G}_1 X E E^\top X^\top \mathcal{G}_1^2\mathbf{u}_1\Big)\\
&\quad - \frac{2}{(1-r)^2}r^{-\frac{3}{4}}\mathbf{v}_1^\top X^\top \mathcal{G}_1 X E E^\top X^\top \mathcal{G}_1^2 X E E^\top X^\top \mathcal{G}_1\mathbf{u}_1 + O_\prec(n^{-1}r^{\frac{1}{2}}\|\mathbf{u}_1\|)\,.
\end{aligned}
$$

Further, by (4.9), (4.10), and (D.7), we have

$$
D_3 = \frac{2r^{-\frac{1}{4}}}{(1-r)^2}\mathbf{v}_1^\top X^\top \mathcal{G}_1\mathbf{u}_1 - \frac{2r^{-\frac{3}{4}}}{1-r}\mathbf{v}_1^\top X^\top \mathcal{G}_1^2\mathbf{u}_1 + O_\prec(n^{-1}r^{\frac{1}{2}}\|\mathbf{u}_1\|)\,. \qquad \text{(D.28)}
$$

Combining (D.26), (D.27) and (D.28), we conclude that

$$
\begin{aligned}
&\frac{1}{(1-r)^2}\widehat{A}^\top\widehat{\Sigma}\widehat{A} - \widehat{A}^\top\Sigma\widehat{A}\\
&= \frac{1-2r}{(1-r)^4}\mathbf{v}_1^\top(z\mathcal{G}_2 - zm_2)\mathbf{v}_1 - \frac{r^{-\frac{1}{2}}}{(1-r)^2}\mathbf{v}_1^\top\Big((z\mathcal{G}_2)' - (zm_2)'\Big)\mathbf{v}_1 + \frac{r^{-\frac{1}{2}}}{(1-r)^2}\mathbf{u}_1^\top(\mathcal{G}_1 - m_1)\mathbf{u}_1\\
&\quad - r^{-1}\mathbf{u}_1^\top(\mathcal{G}_1^2 - m_1')\mathbf{u}_1 + \frac{2r^{-\frac{1}{4}}}{(1-r)^2}\mathbf{u}_1^\top\mathcal{G}_1 X\mathbf{v}_1 - \frac{2r^{-\frac{3}{4}}}{1-r}\mathbf{u}_1^\top\mathcal{G}_1^2 X\mathbf{v}_1 + O_\prec\big(n^{-1}(\|\mathbf{u}_1\|^2 + r)\big)\,.
\end{aligned}
$$

(D.29)

Then, expanding $\sqrt{\widehat{A}^\top \Sigma \widehat{A}}$ around $\sqrt{\widehat{A}^\top \widehat{\Sigma} \widehat{A}}/(1-r)$, we finally obtain

$$\begin{aligned}
&\frac{\sqrt{\widehat{A}^\top \widehat{\Sigma} \widehat{A}}}{1-r} - \sqrt{\widehat{A}^\top \Sigma \widehat{A}} = \frac{1-r}{2\sqrt{\widehat{A}^\top \widehat{\Sigma} \widehat{A}}}\bigg(\frac{1-2r}{(1-r)^4} \mathbf{v}_1^\top \big(z\mathcal{G}_2 - zm_2\big)\mathbf{v}_1 - \frac{r^{-\frac{1}{2}}}{(1-r)^2}\mathbf{v}_1^\top \Big((z\mathcal{G}_2)' - (zm_2)'\Big)\mathbf{v}_1 \\
&\quad + \frac{r^{-\frac{1}{2}}}{(1-r)^2}\mathbf{u}_1^\top(\mathcal{G}_1 - m_1)\mathbf{u}_1 - r^{-1}\mathbf{u}_1^\top(\mathcal{G}_1^2 - m_1')\mathbf{u}_1 + \frac{2r^{-\frac{1}{4}}}{(1-r)^2}\mathbf{u}_1^\top \mathcal{G}_1 X\mathbf{v}_1 - \frac{2r^{-\frac{3}{4}}}{1-r}\mathbf{u}_1^\top \mathcal{G}_1^2 X\mathbf{v}_1\bigg) \\
&\quad + O_\prec\big(n^{-1}(r^{\frac{1}{2}} + \|\mathbf{u}_1\|)\big)\,.
\end{aligned}$$

Next, analogously, we have

$$\begin{aligned}
&\widehat{A}^\top \hat{\boldsymbol{\mu}}^0 - \sqrt{\frac{n}{n_0}}\frac{r}{1-r}\mathbf{v}_1^\top \mathbf{e}_0 - \widehat{A}^\top \boldsymbol{\mu}^0 \\
&= \frac{n-2}{\sqrt{nn_0}}\Big(\mathbf{v}_1^\top X^\top \mathcal{G}_1 X\mathbf{e}_0 + \mathbf{v}_1^\top X^\top \mathcal{G}_1 X E\mathcal{I}_2^{-1}E^\top X^\top \mathcal{G}_1 X\mathbf{e}_0\Big) \\
&\quad + \frac{n-2}{\sqrt{nn_0}}r^{-\frac{1}{4}}\Big(\mathbf{u}_1^\top \mathcal{G}_1 X\mathbf{e}_0 + \mathbf{u}_1^\top \mathcal{G}_1 X E\mathcal{I}_2^{-1}E^\top X^\top \mathcal{G}_1 X\mathbf{e}_0\Big) - \sqrt{\frac{n}{n_0}}\frac{r}{1-r}\mathbf{v}_1^\top \mathbf{e}_0 \\
&= \sqrt{\frac{n}{n_0}}\Big(\mathbf{v}_1^\top(z\mathcal{G}_2 - zm_2)\mathbf{e}_0 + \mathbf{v}_1^\top X^\top \mathcal{G}_1 X E\mathcal{I}_2^{-1}E^\top X^\top \mathcal{G}_1 X\mathbf{e}_0 - \frac{r^2}{1-r}\mathbf{v}_1^\top \mathbf{e}_0\Big) \\
&\quad + \sqrt{\frac{n}{n_0}}r^{-\frac{1}{4}}\Big(\mathbf{u}_1^\top \mathcal{G}_1 X\mathbf{e}_0 + \mathbf{u}_1^\top \mathcal{G}_1 X E\mathcal{I}_2^{-1}E^\top X^\top \mathcal{G}_1 X\mathbf{e}_0\Big) + O_\prec(n^{-\frac{1}{2}}n_0^{-\frac{1}{2}}r^{\frac{1}{2}}(r^{\frac{1}{2}} + \|\mathbf{u}_1\|))\,.
\end{aligned}$$

Employing (D.25) and the estimates (4.9), (4.10) with (D.7), we can further get that

$$\begin{aligned}
&\widehat{A}^\top \hat{\boldsymbol{\mu}}^0 - \sqrt{\frac{n}{n_0}}\frac{r}{1-r}\mathbf{v}_1^\top \mathbf{e}_0 - \widehat{A}^\top \boldsymbol{\mu}^0 \\
&= \sqrt{\frac{n}{n_0}}\Big(\frac{1}{(1-r)^2}\mathbf{v}_1^\top(z\mathcal{G}_2 - zm_2)\mathbf{e}_0 + \frac{r^{-\frac{1}{4}}}{1-r}\mathbf{u}_1^\top \mathcal{G}_1 X\mathbf{e}_0\Big) + O_\prec\big(n^{-\frac{1}{2}}n_0^{-\frac{1}{2}}r^{\frac{1}{2}}(r^{\frac{1}{2}} + \|\mathbf{u}_1\|)\big)\,.
\end{aligned} \tag{D.30}$$

In light of (D.30) and (D.29), together with the fact $\|\mathbf{u}_1\|^2 = \mathit{\Delta}_d$, we can now conclude the proof of Lemma D.1. ■

In the sequel, we state the proof of Proposition D.1 which will rely on Gaussian integration by parts. For simplicity, we always drop z-dependence from the notations $\mathcal{G}_1(z)$, $\mathcal{G}_2(z)$ and $m_1(z), m_2(z)$. We also fix the choice of z in (D.22) and omit this fact from the notations.

Recall the definition of $\mathcal{P}$ in (D.21). For brevity, we introduce the shorthand notation

$$\begin{aligned}
&\mathbf{y}_0 := c_{10}\bar{\mathbf{u}}_1, \quad \mathbf{y}_1 := c_{11}\bar{\mathbf{u}}_1, \quad \tilde{\mathbf{y}}_0 := c_{12}\bar{\mathbf{v}}_1 + c_{13}\mathbf{e}_0, \quad \tilde{\mathbf{y}}_1 := c_{14}\bar{\mathbf{v}}_1, \\
&\boldsymbol{\eta}_0 := c_{20}\bar{\mathbf{v}}_1 + c_{21}\mathbf{e}_0, \quad \boldsymbol{\eta}_1 := c_{22}\bar{\mathbf{v}}_1\,.
\end{aligned} \tag{D.31}$$

Using the basic identity $z\mathcal{G}_2 = X^\top \mathcal{G}_1 X - I_n$, we can simplify the expression of $\mathcal{P}$ in (D.21) to

$$\begin{aligned}\mathcal{P} =& \sqrt{n}\sum_{t=0}^{1}\Big(r^{-\frac{1+t}{2}}(\mathcal{G}_1^{(t)} - m_1^{(t)})_{\bar{\mathbf{u}}_1\mathbf{y}_t} + r^{-\frac{1+2t}{4}}(\mathcal{G}_1^{(t)}X)_{\bar{\mathbf{u}}_1\tilde{\mathbf{y}}_t} \\ &\qquad + r^{-\frac{1+t}{2}}\Big((X^\top\mathcal{G}_1X)^{(t)} - (1+zm_2)^{(t)}\Big)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_t}\Big)\,. \end{aligned} \tag{D.32}$$

Further, by Proposition 1 and Remark 2, it is easy to see

$$\mathcal{P} = O_{\prec}(1)\,. \tag{D.33}$$

Using the identity

$$\mathcal{G}_1^t = z^{-1}(H\mathcal{G}_1^t - \mathcal{G}_1^{t-1})\,, \qquad t = 1, 2\,,$$

we can rewrite

$$\begin{aligned}&\sqrt{n}\sum_{t=0,1} r^{-\frac{1+t}{2}}(\mathcal{G}_1^{(t)} - m_1^{(t)})_{\bar{\mathbf{u}}_1\mathbf{y}_t} \\ &= \frac{\sqrt{n}}{r}\Bigg(\frac{1}{(1+r^{-\frac{1}{2}}m_2)z}(H\mathcal{G}_1^2)_{\bar{\mathbf{u}}_1\mathbf{y}_1} + \frac{r^{-\frac{1}{2}}m_2}{1+r^{-\frac{1}{2}}m_2}(\mathcal{G}_1^2)_{\bar{\mathbf{u}}_1\mathbf{y}_1} \\ &\qquad + \Big(\frac{r^{-\frac{1}{2}}(zm_2)'}{(1+r^{-\frac{1}{2}}m_2)z} - \frac{r^{-\frac{1}{2}}(zm_2)'+1}{(1+r^{-\frac{1}{2}}m_2)z}\Big)(\mathcal{G}_1)_{\bar{\mathbf{u}}_1\mathbf{y}_1} - m_1'\big(\bar{\mathbf{u}}_1\big)^\top\mathbf{y}_1\Bigg) \\ &\quad + \frac{\sqrt{n}}{\sqrt{r}}\Bigg(\frac{1}{(1+r^{-\frac{1}{2}}m_2)z}(H\mathcal{G}_1)_{\bar{\mathbf{u}}_1\mathbf{y}_0} + \frac{r^{-\frac{1}{2}}m_2}{1+r^{-\frac{1}{2}}m_2}(\mathcal{G}_1)_{\bar{\mathbf{u}}_1\mathbf{y}_0} - \frac{1}{(1+r^{-\frac{1}{2}}m_2)z}\big(\bar{\mathbf{u}}_1\big)^\top\mathbf{y}_0 - m_1\big(\bar{\mathbf{u}}_1\big)^\top\mathbf{y}_0\Bigg) \\ &= \frac{\sqrt{n}}{r}\Bigg(\frac{1}{(1+r^{-\frac{1}{2}}m_2)z}(H\mathcal{G}_1^2)_{\bar{\mathbf{u}}_1\mathbf{y}_1} + \frac{r^{-\frac{1}{2}}m_2}{1+r^{-\frac{1}{2}}m_2}(\mathcal{G}_1^2)_{\bar{\mathbf{u}}_1\mathbf{y}_1} + \frac{r^{-\frac{1}{2}}(zm_2)'}{(1+r^{-\frac{1}{2}}m_2)z}(\mathcal{G}_1)_{\bar{\mathbf{u}}_1\mathbf{y}_1}\Bigg) \\ &\quad + \frac{m_1'}{m_1}\frac{\sqrt{n}}{r}\Bigg(\frac{1}{(1+r^{-\frac{1}{2}}m_2)z}(H\mathcal{G}_1)_{\bar{\mathbf{u}}_1\mathbf{y}_1} + \frac{r^{-\frac{1}{2}}m_2}{1+r^{-\frac{1}{2}}m_2}(\mathcal{G}_1)_{\bar{\mathbf{u}}_1\mathbf{y}_1}\Bigg) \\ &\quad + \frac{\sqrt{n}}{\sqrt{r}}\Bigg(\frac{1}{(1+r^{-\frac{1}{2}}m_2)z}(H\mathcal{G}_1)_{\bar{\mathbf{u}}_1\mathbf{y}_0} + \frac{r^{-\frac{1}{2}}m_2}{1+r^{-\frac{1}{2}}m_2}(\mathcal{G}_1)_{\bar{\mathbf{u}}_1\mathbf{y}_0}\Bigg) \\ &=: \mathbb{T}_{11} + \mathbb{T}_{12} + \mathbb{T}_{13}\,. \end{aligned} \tag{D.34}$$

Here we used the first and last identities in (D.1) to gain some cancellations. Particularly, from first step to second step, we also do the derivation

$$\begin{aligned}-\frac{r^{-\frac{1}{2}}(zm_2)'+1}{(1+r^{-\frac{1}{2}}m_2)z}(\mathcal{G}_1)_{\bar{\mathbf{u}}_1\mathbf{y}_1} - m_1'\big(\bar{\mathbf{u}}_1\big)^\top\mathbf{y}_1 &= \frac{m_1'}{m_1}\Big((\mathcal{G}_1)_{\bar{\mathbf{u}}_1\mathbf{y}_1} - m_1\big(\bar{\mathbf{u}}_1\big)^\top\mathbf{y}_1\Big) \\ &= \frac{m_1'}{m_1}\Bigg(\frac{1}{(1+r^{-\frac{1}{2}}m_2)z}(H\mathcal{G}_1)_{\bar{\mathbf{u}}_1\mathbf{y}_1} + \frac{r^{-\frac{1}{2}}m_2}{1+r^{-\frac{1}{2}}m_2}(\mathcal{G}_1)_{\bar{\mathbf{u}}_1\mathbf{y}_1}\Bigg)\,. \end{aligned}$$

Next, we also rewrite

$$
\begin{aligned}
&\sqrt{n}\sum_{t=0}^{1} r^{-\frac{1+2t}{4}}(\mathcal{G}_1^{(t)}X)_{\bar{\mathbf{u}}_1\tilde{\mathbf{y}}_t}\\
&=\sqrt{n}r^{-\frac{3}{4}}\bigg((\mathcal{G}_1^2X)_{\bar{\mathbf{u}}_1\tilde{\mathbf{y}}_1}+\frac{r^{\frac{1}{2}}m_1'}{1+r^{\frac{1}{2}}m_1}(\mathcal{G}_1X)_{\bar{\mathbf{u}}_1\tilde{\mathbf{y}}_1}\bigg)-\frac{r^{-\frac{1}{4}}m_1'}{1+r^{\frac{1}{2}}m_1}\sqrt{n}(\mathcal{G}_1X)_{\bar{\mathbf{u}}_1\tilde{\mathbf{y}}_1}+\sqrt{n}r^{-\frac{1}{4}}(\mathcal{G}_1X)_{\bar{\mathbf{u}}_1\tilde{\mathbf{y}}_0}\\
&=:\mathbb{T}_{21}+\mathbb{T}_{22}+\mathbb{T}_{23}\,, \qquad\qquad (\text{D.35})
\end{aligned}
$$

and

$$
\begin{aligned}
&\sqrt{n}\sum_{t=0}^{1} r^{-\frac{1+t}{2}}\Big((X^\top\mathcal{G}_1X)^{(t)}-(1+zm_2)^{(t)}\Big)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_t}\\
&=\frac{\sqrt{n}}{r}\bigg((X^\top\mathcal{G}_1^2X)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_1}+\frac{\sqrt{r}m_1'}{1+\sqrt{r}m_1}(X^\top\mathcal{G}_1X)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_1}-\frac{\sqrt{r}m_1'}{1+\sqrt{r}m_1}\big(\bar{\mathbf{v}}_1\big)^\top\boldsymbol{\eta}_1\bigg)\\
&\quad-\frac{\sqrt{r}m_1'}{1+\sqrt{r}m_1}\frac{\sqrt{n}}{r}\bigg((X^\top\mathcal{G}_1X)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_1}-\Big(1-\frac{1+\sqrt{r}m_1}{\sqrt{r}m_1'}(zm_2)'\Big)\big(\bar{\mathbf{v}}_1\big)^\top\boldsymbol{\eta}_1\bigg)\\
&\quad+\frac{\sqrt{n}}{\sqrt{r}}\bigg((X^\top\mathcal{G}_1X)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_0}-(1+zm_2)\big(\bar{\mathbf{v}}_1\big)^\top\boldsymbol{\eta}_0\bigg)\\
&=\frac{\sqrt{n}}{r}\bigg((X^\top\mathcal{G}_1^2X)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_1}+\frac{\sqrt{r}m_1'}{1+\sqrt{r}m_1}(z\mathcal{G}_2)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_1}\bigg)\\
&\quad-\frac{\sqrt{r}m_1'}{1+\sqrt{r}m_1}\frac{\sqrt{n}}{r}\bigg(\frac{1}{1+\sqrt{r}m_1}(X^\top\mathcal{G}_1X)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_1}+\frac{\sqrt{r}m_1}{1+\sqrt{r}m_1}(z\mathcal{G}_2)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_1}\bigg)\\
&\quad+\frac{\sqrt{n}}{\sqrt{r}}\bigg(\frac{1}{1+\sqrt{r}m_1}(X^\top\mathcal{G}_1X)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_0}+\frac{\sqrt{r}m_1}{1+\sqrt{r}m_1}(z\mathcal{G}_2)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_0}\bigg)\\
&=:\mathbb{T}_{31}+\mathbb{T}_{32}+\mathbb{T}_{33}\,, \qquad\qquad (\text{D.36})
\end{aligned}
$$

where we used the second identity in (D.5) and the identities

$$
1-\frac{1+\sqrt{r}m_1}{\sqrt{r}m_1'}(zm_2)'=1+zm_2\,,\quad \frac{\sqrt{r}m_1}{1+\sqrt{r}m_1}=1+zm_2\,. \qquad (\text{D.37})
$$

We remark here that (D.37) can be easily checked by applying the identities in (D.1), the first equation in (4.3), and also the identity obtained by taking derivative w.r.t z for both sides of the first equation in (4.3), i.e.,

$$
\sqrt{r}m_1^2+2z\sqrt{r}\,m_1m_1'+m_1+(z-1/\sqrt{r}+\sqrt{r})m_1'=0\,.
$$

Before we commence the proof of Proposition D.1, let us first state below the derivative of $\mathcal{P}$, which follows from a direct calculation

$$\begin{aligned}\frac{\partial \mathcal{P}}{\partial x_{ij}} = &-\sqrt{n} \sum_{\substack{a_1,a_2\geq 1\\ a=a_1+a_2\leq 3}} r^{-\frac{a-1}{2}} \Big((\mathcal{G}_1^{a_1}\bar{\mathbf{u}}_1)_i (X^\top \mathcal{G}_1^{a_2}\mathbf{y}_{a-2})_j + (X^\top \mathcal{G}_1^{a_1}\bar{\mathbf{u}}_1)_j (\mathcal{G}_1^{a_2}\mathbf{y}_{a-2})_i \Big)\\ &- \sqrt{n} \sum_{\substack{a_1,a_2\geq 1\\ a=a_1+a_2\leq 3}} r^{-\frac{1+2(a-2)}{4}} \Big((\mathcal{G}_1^{a_1}\bar{\mathbf{u}}_1)_i ((z\mathcal{G}_2)^{(a_2-1)}\tilde{\mathbf{y}}_{a_2})_j + (X^\top \mathcal{G}_1^{a_1}\bar{\mathbf{u}}_1)_j (\mathcal{G}_1^{a_2}X\tilde{\mathbf{y}}_{a-2})_i \Big)\\ &- \sqrt{n} \sum_{\substack{a_1,a_2\geq 1\\ a=a_1+a_2\leq 3}} r^{-\frac{a-1}{2}} \Big((\mathcal{G}_1^{a_1}X\bar{\mathbf{v}}_1)_i \big((z\mathcal{G}_2)^{(a_2-1)}\boldsymbol{\eta}_{a-2}\big)_j + (\mathcal{G}_1^{a_1}X\boldsymbol{\eta}_{a-2})_i \big((z\mathcal{G}_2)^{(a_2-1)}\bar{\mathbf{v}}_1\big)_j \Big).\end{aligned} \tag{D.38}$$

Now, let us proceed to the proof of Proposition D.1.

Proof (Proof of Proposition D.1)

By the definition of characteristic function, we have, for $t \in \mathbf{R}$,

$$\varphi_n(t) = \mathbb{E}e^{\mathrm{i}t\mathcal{P}}\,, \quad \varphi_n'(t) = \mathrm{i}\mathbb{E}\mathcal{P}e^{\mathrm{i}t\mathcal{P}}\,.$$

We will estimate $\varphi_n'(t)$ via Gaussian integration by parts. Recall the representation of $\mathcal{P}$ in (D.32) together with (D.34)-(D.36), we may further express

$$\varphi_n'(t) = \mathrm{i}\mathbb{E}\sum_{i,j=1}^{3} \mathbb{T}_{ij} h(t)\,, \qquad h(t) := e^{\mathrm{i}t\mathcal{P}}\,.$$

For convenience, we use the following shorthand notation for summation

$$\sum_{i,j} := \sum_{i=1}^{p}\sum_{j=1}^{n}\,.$$

Since all entries x_{ij} are i.i.d $\mathcal{N}(0, 1/\sqrt{np})$, applying Gaussian integration by parts leads to

$$\begin{aligned}\mathrm{i}\mathbb{E}\frac{\sqrt{n}}{r}(H\mathcal{G}_1^2)_{\bar{\mathbf{u}}_1\mathbf{y}_1}\, h(t) &= \mathrm{i}\frac{\sqrt{n}}{r}\sum_{i,j}\mathbb{E}\bar{u}_{1i}x_{ij}(X^\top\mathcal{G}_1^2\mathbf{y}_1)_j\, h(t) = \mathrm{i}\frac{r^{-\frac{3}{2}}}{\sqrt{n}}\,\mathbb{E}\sum_{i,j}\bar{u}_{1i}\frac{\partial(X^\top\mathcal{G}_1^2\mathbf{y}_1)_j h(t)}{\partial x_{ij}}\\ &= \mathrm{i}\mathbb{E}\Big(\sqrt{n}r^{-\frac{3}{2}}(\mathcal{G}_1^2)_{\bar{\mathbf{u}}_1\mathbf{y}_1} - \frac{r^{-\frac{3}{2}}}{\sqrt{n}}(\mathcal{G}_1 H\mathcal{G}_1^2)_{\bar{\mathbf{u}}_1\mathbf{y}_1} - \frac{r^{-\frac{3}{2}}}{\sqrt{n}}(\mathcal{G}_1^2 H\mathcal{G}_1)_{\bar{\mathbf{u}}_1\mathbf{y}_1}\\ &\qquad - \sqrt{n}r^{-\frac{3}{2}}(\mathcal{G}_1^2)_{\bar{\mathbf{u}}_1\mathbf{y}_1}\frac{\operatorname{Tr} X^\top\mathcal{G}_1X}{n} - \sqrt{n}r^{-\frac{3}{2}}(\mathcal{G}_1)_{\bar{\mathbf{u}}_1\mathbf{y}_1}\frac{\operatorname{Tr} X^\top\mathcal{G}_1^2X}{n}\Big)h(t)\\ &\quad + \frac{\mathrm{i}^2 t}{\sqrt{nr^3}}\mathbb{E}\sum_{i,j}\bar{u}_{1i}(X^\top\mathcal{G}_1^2\mathbf{y}_1)_j\frac{\partial\mathcal{P}}{\partial x_{ij}}h(t)\,.\end{aligned}$$

Then, by Proposition 1, Remark 2, and the fact $m_1^{(a)}(z) = O(r^{(1+a)/2})$ for $a = 0, 1, 2$ owing to the choice of z in (D.22), we further have

$$\begin{aligned}
\mathrm{i}\mathbb{E}\frac{\sqrt{n}}{r}(H\mathcal{G}_1^2)_{\bar{\mathbf{u}}_1\mathbf{y}_1}h(t) &= \mathrm{i}\mathbb{E}\Big(-\sqrt{n}r^{-\frac{3}{2}}(\mathcal{G}_1^2)_{\bar{\mathbf{u}}_1\mathbf{y}_1}\frac{\operatorname{Tr} z\mathcal{G}_2}{n} - \sqrt{n}r^{-\frac{3}{2}}(\mathcal{G}_1)_{\bar{\mathbf{u}}_1\mathbf{y}_1}\frac{\operatorname{Tr}(z\mathcal{G}_2)'}{n} \\
&\qquad - n^{-\frac{1}{2}}r^{-\frac{3}{2}}(z\mathcal{G}_1)''_{\bar{\mathbf{u}}_1\mathbf{y}_1}\Big)h(t) + \frac{\mathrm{i}^2 t}{\sqrt{nr^3}}\mathbb{E}\sum_{i,j}\bar{u}_{1i}(X^\top\mathcal{G}_1^2\mathbf{y}_1)_j\frac{\partial\mathcal{P}}{\partial x_{ij}}h(t) \\
&= -\mathrm{i}\sqrt{n}r^{-\frac{3}{2}}\mathbb{E}\Big(zm_2(\mathcal{G}_1^2)_{\bar{\mathbf{u}}_1\mathbf{y}_1} + (zm_2)'(\mathcal{G}_1)_{\bar{\mathbf{u}}_1\mathbf{y}_1}\Big)h(t) \\
&\qquad + \frac{\mathrm{i}^2 t}{\sqrt{nr^3}}\mathbb{E}\sum_{i,j}\bar{u}_{1i}(X^\top\mathcal{G}_1^2\mathbf{y}_1)_j\frac{\partial\mathcal{P}}{\partial x_{ij}}h(t) + O_\prec(p^{-\frac{1}{2}}) . \qquad \text{(D.39)}
\end{aligned}$$

Next, plugging in (D.38), we have the term

$$\begin{aligned}
&\frac{\mathrm{i}^2 t}{\sqrt{nr^3}}\mathbb{E}\sum_{i,j}\bar{u}_{1i}(X^\top\mathcal{G}_1^2\mathbf{y}_1)_j\frac{\partial\mathcal{P}}{\partial x_{ij}}h(t) \\
&= -\mathrm{i}^2 tr^{-\frac{3}{2}}\mathbb{E}\sum_{i,j}\bar{u}_{1i}(X^\top\mathcal{G}_1^2\mathbf{y}_1)_j\sum_{\substack{a_1,a_2\geq 1\\ a=a_1+a_2\leq 3}}\Big[r^{-\frac{a-1}{2}}\Big((\mathcal{G}_1^{a_1}\bar{\mathbf{u}}_1)_i(X^\top\mathcal{G}_1^{a_2}\mathbf{y}_{a-2})_j + (X^\top\mathcal{G}_1^{a_1}\bar{\mathbf{u}}_1)_j(\mathcal{G}_1^{a_2}\mathbf{y}_{a-2})_i\Big) \\
&\quad + r^{-\frac{1+2(a-2)}{4}}\Big((\mathcal{G}_1^{a_1}\bar{\mathbf{u}}_1)_i\big((z\mathcal{G}_2)^{(a_2-1)}\tilde{\mathbf{y}}_{a-2}\big)_j + (X^\top\mathcal{G}_1^{a_1}\bar{\mathbf{u}}_1)_j(\mathcal{G}_1^{a_2}X\tilde{\mathbf{y}}_{a-2})_i\Big) \\
&\quad + r^{-\frac{a-1}{2}}\Big((\mathcal{G}_1^{a_1}X\boldsymbol{\eta}_{a-2})_i\big((z\mathcal{G}_2)^{(a_2-1)}\bar{\mathbf{v}}_1\big)_j + (\mathcal{G}_1^{a_1}X\bar{\mathbf{v}}_1)_i\big((z\mathcal{G}_2)^{(a_2-1)}\boldsymbol{\eta}_{a-2}\big)_j\Big)\Big]h(t) . \qquad \text{(D.40)}
\end{aligned}$$

It is easy to see that the RHS of the above equation is a linear combination of the expectations of the terms taking the following forms

$$\begin{aligned}
&tr^{-\frac{b_1+b_2+b_3}{2}}\big(\vartheta_1^\top\mathcal{G}_1^{b_1}\vartheta_2\big)\big(\vartheta_3^\top\mathcal{G}_1^{b_2}H\mathcal{G}_1^{b_3}\vartheta_4\big) , \qquad tr^{-\frac{2(b_1+b_2+b_3)-1}{4}}\big(\vartheta_1^\top\mathcal{G}_1^{b_1}\vartheta_2\big)\big(\vartheta_3^\top\mathcal{G}_1^{b_2}X(z\mathcal{G}_2)^{(b_3-1)}\vartheta_4\big) , \\
&tr^{-\frac{2(b_1+b_2+b_3)-1}{4}}\big(\vartheta_1^\top\mathcal{G}_1^{b_1}X\vartheta_2\big)\big(\vartheta_3^\top\mathcal{G}_1^{b_2}H\mathcal{G}_1^{b_3}\vartheta_4\big) , \qquad tr^{-\frac{b_1+b_2+b_3}{2}}\big(\vartheta_1^\top\mathcal{G}_1^{b_1}X\vartheta_2\big)\big(\vartheta_3^\top\mathcal{G}_1^{b_2}X(z\mathcal{G}_2)^{(b_3-1)}\vartheta_4\big) . \\
&\qquad \text{(D.41)}
\end{aligned}$$

Here ϑ_i, $i = 1, 2, 3, 4$ represent for vectors of suitable dimension and $b_i = 1, 2$, for $i = 1, 2, 3$. By (4.9), (4.10) and the fact $m_1^{(a)}(z) = O(r^{(1+a)/2})$ for $a \in \mathbb{N}$, it is easy to observe that except for the first term in (D.41), all the others can be bounded by $O_\prec(tp^{-1/2})$. For instance, for the factor $\big(\vartheta_3^\top\mathcal{G}_1^{b_2}X(z\mathcal{G}_2)^{(b_3-1)}\vartheta_4\big)$, we can use the following estimates which are consequences of (4.10),

$$\begin{aligned}
&\vartheta_3^\top\mathcal{G}_1^{b_2}X(z\mathcal{G}_2)\vartheta_4 = \vartheta_3^\top z\mathcal{G}_1^{b_2+1}X\vartheta_4 = O_\prec(n^{-\frac{1}{2}}r^{\frac{1+2b_2}{4}}) , \\
&\vartheta_3^\top\mathcal{G}_1^{b_2}X(z\mathcal{G}_2)'\vartheta_4 = \vartheta_3^\top\mathcal{G}_1^{b_2}XX^\top\mathcal{G}_1^2X\vartheta_4 = \vartheta_3^\top\mathcal{G}_1^{b_2+1}X\vartheta_4 + z\vartheta_3^\top\mathcal{G}_1^{b_2+2}X\vartheta_4 = O_\prec(n^{-\frac{1}{2}}r^{\frac{1+2b_2}{4}}) .
\end{aligned}$$

Therefore, by the above discussion, we can further simplify (D.40) to get

$$
\begin{aligned}
&\frac{\mathrm{i}^2 t}{\sqrt{nr^3}}\mathbb{E}\sum_{i,j}\bar{u}_{1i}(X^\top\mathcal{G}_1^2\mathbf{y}_1)_j\frac{\partial\mathcal{P}}{\partial x_{ij}}h(t)\\
&=-\mathrm{i}^2 t\mathbb{E}\sum_{\substack{a_1,a_2\geq 1\\ a=a_1+a_2\leq 3}} r^{-\frac{a+2}{2}}\Big((\mathcal{G}_1^{a_1})_{\bar{\mathbf{u}}_1\bar{\mathbf{u}}_1}(\mathcal{G}_1^2H\mathcal{G}_1^{a_2})_{\mathbf{y}_1\mathbf{y}_{a-2}}+(\mathcal{G}_1^{a_2})_{\bar{\mathbf{u}}_1\mathbf{y}_{a-2}}(\mathcal{G}_1^2H\mathcal{G}_1^{a_1})_{\mathbf{y}_1\bar{\mathbf{u}}_1}\Big)h(t)+O_\prec(|t|p^{-\frac{1}{2}})\\
&=-\mathrm{i}^2 t\sum_{\substack{a_1,a_2\geq 1\\ a=a_1+a_2\leq 3}} r^{-\frac{a+2}{2}}m_1^{(a_1-1)}\frac{(zm_1)^{(a_2+1)}}{(a_2+1)!}\Big(\mathbf{y}_1^\top\mathbf{y}_{a-2}+\big(\bar{\mathbf{u}}_1\big)^\top\mathbf{y}_{a-2}\,\mathbf{y}_1^\top\bar{\mathbf{u}}_1\Big)\varphi_n(t)+O_\prec(|t|p^{-\frac{1}{2}})\\
&=t\Big[\frac{m_1(zm_1)''}{2r^2}\Big(\mathbf{y}_1^\top\mathbf{y}_0+\big(\bar{\mathbf{u}}_1\big)^\top\mathbf{y}_0\,\mathbf{y}_1^\top\bar{\mathbf{u}}_1\Big)+\Big(\frac{m_1'(zm_1)''}{2r^{\frac{5}{2}}}+\frac{m_1(zm_1)'''}{3!r^{\frac{5}{2}}}\Big)\Big(\|\mathbf{y}_1\|^2+(\mathbf{y}_1^\top\bar{\mathbf{u}}_1)^2\Big)\Big]\varphi_n(t)\\
&\quad+O_\prec(|t|p^{-\frac{1}{2}})\,. \qquad\qquad (\mathrm{D.42})
\end{aligned}
$$

Combining (D.39) and (D.42), by the definition of $\mathbb{T}_{11}$ in (D.34) and the fact $m_1(z)=O(\sqrt{r})$, we get

$$
\begin{aligned}
\mathrm{i}\mathbb{E}\mathbb{T}_{11}h(t)=-tm_1\Big[&\frac{m_1(zm_1)''}{2r^2}\Big(\mathbf{y}_1^\top\mathbf{y}_0+\big(\bar{\mathbf{u}}_1\big)^\top\mathbf{y}_0\,\mathbf{y}_1^\top\bar{\mathbf{u}}_1\Big)+\Big(\frac{m_1'(zm_1)''}{2r^{\frac{5}{2}}}+\frac{m_1(zm_1)'''}{3!r^{\frac{5}{2}}}\Big)\\
&\times\Big(\|\mathbf{y}_1\|^2+(\mathbf{y}_1^\top\bar{\mathbf{u}}_1)^2\Big)\Big]\varphi_n(t)+O_\prec((|t|+1)n^{-\frac{1}{2}}). \qquad (\mathrm{D.43})
\end{aligned}
$$

By similar arguments, we can also derive

$$
\begin{aligned}
\mathrm{i}\mathbb{E}\mathbb{T}_{12}h(t)=-tm_1'\Big[&r^{-2}m_1(zm_1)'\Big(\mathbf{y}_1^\top\mathbf{y}_0+\big(\bar{\mathbf{u}}_1\big)^\top\mathbf{y}_0\,\mathbf{y}_1^\top\bar{\mathbf{u}}_1\Big)+r^{-\frac{5}{2}}\Big(m_1'(zm_1)'+\frac{m_1(zm_1)''}{2}\Big)\\
&\times\Big(\|\mathbf{y}_1\|^2+(\mathbf{y}_1^\top\bar{\mathbf{u}}_1)^2\Big)\Big]\varphi_n(t)+O_\prec((|t|+1)n^{-\frac{1}{2}})\,, \qquad (\mathrm{D.44})
\end{aligned}
$$

and

$$
\begin{aligned}
\mathrm{i}\mathbb{E}\mathbb{T}_{13}h(t)=-tm_1\Big[&r^{-\frac{3}{2}}m_1(zm_1)'\Big(\|\mathbf{y}_0\|^2+(\mathbf{y}_0^\top\bar{\mathbf{u}}_1)^2\Big)+r^{-2}\Big(m_1'(zm_1)'+\frac{m_1(zm_1)''}{2}\Big)\\
&\times\Big(\mathbf{y}_0^\top\mathbf{y}_1+\big(\bar{\mathbf{u}}_1\big)^\top\mathbf{y}_1\,\mathbf{y}_0^\top\bar{\mathbf{u}}_1\Big)\Big]\varphi_n(t)+O_\prec((|t|+1)n^{-\frac{1}{2}})\,. \qquad (\mathrm{D.45})
\end{aligned}
$$

Next, we turn to study $\mathrm{i}\mathbb{E}\mathbb{T}_{2i}h(t)$, $i=1,2,3$, as defined in (D.35). We first do the decompositions of $\mathbb{T}_{2i}$'s below

$$
\begin{aligned}
\mathbb{T}_{21}&=\sqrt{n}r^{-\frac{3}{4}}\Big(\frac{1}{1+r^{\frac{1}{2}}m_1}(\mathcal{G}_1^2X)_{\bar{\mathbf{u}}_1\tilde{\mathbf{y}}_1}+\frac{r^{\frac{1}{2}}m_1}{1+r^{\frac{1}{2}}m_1}(\mathcal{G}_1^2X)_{\bar{\mathbf{u}}_1\tilde{\mathbf{y}}_1}+\frac{r^{\frac{1}{2}}m_1'}{1+r^{\frac{1}{2}}m_1}(\mathcal{G}_1X)_{\bar{\mathbf{u}}_1\tilde{\mathbf{y}}_1}\Big)\,,\\
&\qquad (\mathrm{D.46})\\
\mathbb{T}_{22}&=-\frac{r^{-\frac{1}{4}}m_1'}{1+r^{\frac{1}{2}}m_1}\sqrt{n}\Big(\frac{1}{1+r^{\frac{1}{2}}m_1}(\mathcal{G}_1X)_{\bar{\mathbf{u}}_1\tilde{\mathbf{y}}_1}+\frac{r^{\frac{1}{2}}m_1}{1+r^{\frac{1}{2}}m_1}(\mathcal{G}_1X)_{\bar{\mathbf{u}}_1\tilde{\mathbf{y}}_1}\Big)\,,\\
\mathbb{T}_{23}&=\sqrt{n}r^{-\frac{1}{4}}\Big(\frac{1}{1+r^{\frac{1}{2}}m_1}(\mathcal{G}_1X)_{\bar{\mathbf{u}}_1\tilde{\mathbf{y}}_0}+\frac{r^{\frac{1}{2}}m_1}{1+r^{\frac{1}{2}}m_1}(\mathcal{G}_1X)_{\bar{\mathbf{u}}_1\tilde{\mathbf{y}}_0}\Big)\,.
\end{aligned}
$$

And we also remark here, these seemingly artificial decompositions, of the form $\mathcal{G}_1 X = s\mathcal{G}_1 X + (1-s)\mathcal{G}_1 X$ for instance, in the terms $\mathbb{T}_{2i}$'s, are used to facilitate our later derivations. More specifically, to prove Proposition D.1, we will derive a self-consistent equation for the characteristic function of $\mathcal{P}$, for which we will need to apply the basic integration by parts formula for Gaussian variables. In the sequel, very often, we will apply the integration by parts to a part such as $s\mathcal{G}_1 X$ and meanwhile keep the other part $(1-s)\mathcal{G}_1 X$ untouched. One will see that applying integration by parts only partially will help us gain some simple algebraic cancellations. Similar decompositions will also appear in the estimates of $\mathbb{T}_{31}$ term.

In the sequel, we only show the details of the estimate for the $\mathbb{T}_{21}$ term. The other two terms can be estimated similarly, and thus we omit the details. By Gaussian integration by parts, we have

$$\begin{aligned}
&\mathrm{i}E\sqrt{n}r^{-\frac{3}{4}}(\mathcal{G}_1^2X)_{\bar{\mathbf{u}}_1\tilde{\mathbf{y}}_1}h(t) = \frac{\mathrm{i}}{\sqrt{n}}r^{-\frac{5}{4}}\mathbb{E}\sum_{i,j}\tilde{y}_{1j}\frac{\partial(\mathcal{G}_1^2\bar{\mathbf{u}}_1)_i h(t)}{\partial x_{ij}}\\
&= -\mathrm{i}\mathbb{E}\Big(\frac{2r^{-\frac{5}{4}}}{\sqrt{n}}(\mathcal{G}_1^3X)_{\bar{\mathbf{u}}_1\tilde{\mathbf{y}}_1} + \sqrt{n}r^{-\frac{1}{4}}(\mathcal{G}_1^2X)_{\bar{\mathbf{u}}_1\tilde{\mathbf{y}}_1}\frac{\mathrm{Tr}\mathcal{G}_1}{p} + \sqrt{n}r^{-\frac{1}{4}}(\mathcal{G}_1X)_{\bar{\mathbf{u}}_1\tilde{\mathbf{y}}_1}\frac{\mathrm{Tr}\mathcal{G}_1^2}{p}\Big)h(t)\\
&\quad + \frac{\mathrm{i}^2t}{\sqrt{n}}r^{-\frac{5}{4}}\mathbb{E}\sum_{i,j}\tilde{y}_{1j}(\mathcal{G}_1^2\bar{\mathbf{u}}_1)_i\frac{\partial\mathcal{P}}{\partial x_{ij}}h(t)\\
&= -\mathrm{i}\mathbb{E}\Big(\sqrt{n}r^{-\frac{1}{4}}m_1(\mathcal{G}_1^2X)_{\bar{\mathbf{u}}_1\tilde{\mathbf{y}}_1} + \sqrt{n}r^{-\frac{1}{4}}m_1'(\mathcal{G}_1X)_{\bar{\mathbf{u}}_1\tilde{\mathbf{y}}_1}\Big)h(t)\\
&\quad + \frac{\mathrm{i}^2t}{\sqrt{n}}r^{-\frac{5}{4}}\mathbb{E}\sum_{i,j}\tilde{y}_{1j}(\mathcal{G}_1^2\bar{\mathbf{u}}_1)_i\frac{\partial\mathcal{P}}{\partial x_{ij}}h(t) + O_\prec(n^{-\frac{1}{2}}),
\end{aligned}$$

where in the last step we used (4.10), (4.7) and the fact $m_1^{(a)}(z) = O(r^{(1+a)/2})$ for $a = 0, 1$. Plugging the above estimate into the first term in $\mathrm{i}\mathbb{E}\mathbb{T}_{21}h(t)$ which corresponds to the the first term inside the parenthesis in (D.46), we easily see that

$$\mathrm{i}\mathbb{E}\mathbb{T}_{21}h(t) = \frac{r^{-\frac{5}{4}}}{1+\sqrt{r}m_1}\frac{\mathrm{i}^2t}{\sqrt{n}}\mathbb{E}\sum_{i,j}\tilde{y}_{1j}(\mathcal{G}_1^2\bar{\mathbf{u}}_1)_i\frac{\partial\mathcal{P}}{\partial x_{ij}}h(t) + O_\prec(n^{-\frac{1}{2}}).$$

Similarly to (D.40)- (D.42), we can also derive that

$$\begin{aligned}
&\frac{r^{-\frac{5}{4}}}{\sqrt{n}}t\mathbb{E}\sum_{i,j}\tilde{y}_{1j}(\mathcal{G}_1^2\bar{\mathbf{u}}_1)_i\frac{\partial\mathcal{P}}{\partial x_{ij}}h(t)\\
&= -t\mathbb{E}\sum_{i,j}\tilde{y}_{1j}(\mathcal{G}_1^2\bar{\mathbf{u}}_1)_i\sum_{\substack{a_1,a_2\geq 1\\ a=a_1+a_2\leq 3}}r^{-\frac{a+1}{2}}(\mathcal{G}_1^{a_1}\bar{\mathbf{u}}_1)_i\big((z\mathcal{G}_2)^{(a_2-1)}\tilde{\mathbf{y}}_{a-2}\big)_jh(t) + O_\prec(|t|n^{-\frac{1}{2}})\\
&= -t\Big[\frac{m_1''}{2r^{\frac{3}{2}}}(zm_2)\tilde{\mathbf{y}}_1^\top\tilde{\mathbf{y}}_0 + r^{-2}\Big(\frac{m_1'''}{3!}(zm_2) + \frac{m_1''}{2}(zm_2)'\Big)\|\tilde{\mathbf{y}}_1\|^2\Big]\varphi_n(t) + O_\prec(|t|n^{-\frac{1}{2}}), \quad \text{(D.47)}
\end{aligned}$$

which leads to

$$\begin{aligned}
\mathrm{i}\mathbb{E}\mathbb{T}_{21}h(t) = \frac{t}{1+\sqrt{r}m_1}&\Big[\frac{m_1''}{2r^{\frac{3}{2}}}(zm_2)\tilde{\mathbf{y}}_1^\top\tilde{\mathbf{y}}_0 + r^{-2}\Big(\frac{m_1'''}{3!}(zm_2) + \frac{m_1''}{2}(zm_2)'\Big)\|\tilde{\mathbf{y}}_1\|^2\Big]\varphi_n(t)\\
&+ O_\prec((|t|+1)n^{-\frac{1}{2}}). \qquad \text{(D.48)}
\end{aligned}$$

By analogous derivations, we can get the following estimates

$$
\begin{aligned}
\mathrm{i}\mathbb{E}\mathbb{T}_{22}h(t) &= -\frac{tm_1'}{(1+\sqrt{r}m_1)^2}\Big[r^{-1}m_1'(zm_2)\tilde{\mathbf{y}}_1^\top\tilde{\mathbf{y}}_0 + r^{-\frac{3}{2}}\Big(\frac{m_1''}{2}(zm_2) + m_1'(zm_2)'\Big)\|\tilde{\mathbf{y}}_1\|^2\Big]\varphi_n(t) \\
&\quad + O_\prec((|t|+1)n^{-\frac{1}{2}})\,, \qquad \text{(D.49)} \\
\mathrm{i}\mathbb{E}\mathbb{T}_{23}h(t) &= \frac{t}{(1+\sqrt{r}m_1)}\Big[r^{-1}m_1'(zm_2)\|\tilde{\mathbf{y}}_0\|^2 + r^{-\frac{3}{2}}\Big(\frac{m_1''}{2}(zm_2) + m_1'(zm_2)'\Big)\tilde{\mathbf{y}}_0^\top\tilde{\mathbf{y}}_1\Big]\varphi_n(t) \\
&\quad + O_\prec((|t|+1)n^{-\frac{1}{2}})\,. \qquad \text{(D.50)}
\end{aligned}
$$

In the sequel, we focus on the derivation of the estimate of $\mathrm{i}\mathbb{E}\mathbb{T}_{31}h(t)$ and directly conclude the estimates of $\mathrm{i}\mathbb{E}\mathbb{T}_{32}h(t)$, $\mathrm{i}\mathbb{E}\mathbb{T}_{33}h(t)$ without details, since we actually only need to make some adjustments to the estimate of $\mathrm{i}\mathbb{E}\mathbb{T}_{31}h(t)$. First, we do the following artificial decomposition for $\mathbb{T}_{31}$,

$$
\mathbb{T}_{31} = \frac{\sqrt{n}}{r}\Big(\frac{1}{1+\sqrt{r}m_1}(X^\top\mathcal{G}_1^2X)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_1} + \frac{\sqrt{r}m_1}{1+\sqrt{r}m_1}(X^\top\mathcal{G}_1^2X)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_1} + \frac{\sqrt{r}m_1'}{1+\sqrt{r}m_1}(z\mathcal{G}_2)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_1}\Big)\,.
$$

Then, by Gaussian integration by parts, following from (4.10) and (4.7) we have

$$
\begin{aligned}
&\mathrm{i}\mathbb{E}\frac{\sqrt{n}}{r}(X^\top\mathcal{G}^2X)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_1}h(t) \\
&= \mathrm{i}\frac{r^{-\frac{3}{2}}}{\sqrt{n}}\mathbb{E}\sum_{i,j}v_{1j}^0\frac{\partial(\mathcal{G}_1^2X\boldsymbol{\eta}_1)_ih(t)}{\partial x_{ij}} \\
&= -\mathrm{i}\mathbb{E}\Big(\sqrt{n}r^{-\frac{1}{2}}(z\mathcal{G}_2)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_1}\frac{\operatorname{Tr}\mathcal{G}_1^2}{p} + \sqrt{n}r^{-\frac{1}{2}}\big((z\mathcal{G}_2)'\big)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_1}\frac{\operatorname{Tr}\mathcal{G}_1}{p} + \frac{2r^{-\frac{3}{2}}}{\sqrt{n}}(X^\top\mathcal{G}_1^3X)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_1}\Big)h(t) \\
&\quad + \frac{\mathrm{i}^2t}{\sqrt{n}}r^{-\frac{3}{2}}\mathbb{E}\sum_{i,j}v_{1j}^0(\mathcal{G}_1^2X\boldsymbol{\eta}_1)_i\frac{\partial\mathcal{P}}{\partial x_{ij}}h(t) \\
&= -\mathrm{i}\mathbb{E}\Big(\sqrt{n}r^{-\frac{1}{2}}m_1'(z\mathcal{G}_2)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_1} + \sqrt{n}r^{-\frac{1}{2}}m_1\big((z\mathcal{G}_2)'\big)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_1}\Big)h(t) \\
&\quad + \frac{\mathrm{i}^2t}{\sqrt{n}}r^{-\frac{3}{2}}\mathbb{E}\sum_{i,j}v_{1j}^0(\mathcal{G}_1^2X\boldsymbol{\eta}_1)_i\frac{\partial\mathcal{P}}{\partial x_{ij}}h(t) + O_\prec(n^{-\frac{1}{2}})\,.
\end{aligned}
$$

This combined with definition of $\mathbb{T}_{31}$, implies that

$$
\mathrm{i}\mathbb{E}\mathbb{T}_{31}h(t) = -\frac{1}{1+\sqrt{r}m_1}\frac{t}{\sqrt{n}}r^{-\frac{3}{2}}\mathbb{E}\sum_{i,j}v_{1j}^0(\mathcal{G}_1^2X\boldsymbol{\eta}_1)_i\frac{\partial\mathcal{P}}{\partial x_{ij}}h(t) + O_\prec(n^{-\frac{1}{2}})\,.
$$

Referring to (D.47) with slight adjustments, we can easily obtain that

$$\begin{aligned}
&\frac{t}{\sqrt{nr^3}}\mathbb{E}\sum_{i,j} v_{1j}^0(\mathcal{G}_1^2 X\boldsymbol{\eta}_1)_i \frac{\partial \mathcal{P}}{\partial x_{ij}} h(t)\\
&= -t\mathbb{E}\sum_{i,j} v_{1j}^0(\mathcal{G}_1^2 X\boldsymbol{\eta}_1)_i \sum_{\substack{a_1,a_2\geq 1\\ a=a_1+a_2\leq 3}} r^{-\frac{a+2}{2}}\Big((\mathcal{G}_1^{a_1}X\bar{\mathbf{v}}_1)_i\big((z\mathcal{G}_2)^{(a_2-1)}\eta_{a-2}\big)_j + (\mathcal{G}_1^{a_1}X\boldsymbol{\eta}_{a-2})_i\big((z\mathcal{G}_2)^{(a_2-1)}\bar{\mathbf{v}}_1\big)_j\Big)h(t)\\
&\qquad + O_{\prec}(|t|n^{-\frac{1}{2}})\\
&= -t\mathbb{E}\sum_{\substack{a_1,a_2\geq 1\\ a=a_1+a_2\leq 3}} r^{-\frac{a+2}{2}}\Big((X^\top\mathcal{G}_1^{a_1+2}X)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_1}\big((z\mathcal{G}_2)^{(a_2-1)}\big)_{\bar{\mathbf{v}}_1\boldsymbol{\eta}_{a-2}} + (X^\top\mathcal{G}_1^{a_1+2}X)_{\boldsymbol{\eta}_{a-2}\,\boldsymbol{\eta}_1}\big((z\mathcal{G}_2)^{(a_2-1)}\big)_{\bar{\mathbf{v}}_1\bar{\mathbf{v}}_1}\Big)h(t)\\
&\qquad + O_{\prec}(|t|n^{-\frac{1}{2}})\\
&= -t\Big[\frac{(zm_2)''}{2r^2}(zm_2)\Big(\boldsymbol{\eta}_0^\top\boldsymbol{\eta}_1 + \bar{\mathbf{v}}_1^\top\boldsymbol{\eta}_1\,\bar{\mathbf{v}}_1^\top\boldsymbol{\eta}_0\Big)\\
&\qquad + r^{-\frac{5}{2}}\Big(\frac{(zm_2)'''}{3!}(zm_2) + \frac{(zm_2)''}{2}(zm_2)'\Big)\Big(\|\boldsymbol{\eta}_1\|^2 + \big(\bar{\mathbf{v}}_1^\top\boldsymbol{\eta}_1\big)^2\Big)\Big]\varphi_n(t) + O_{\prec}(|t|n^{-\frac{1}{2}})\,.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{i}\mathbb{E}\mathbb{T}_{31}h(t) &= \frac{t}{(1+\sqrt{r}m_1)}\Big[\frac{(zm_2)''}{2r^2}(zm_2)\Big(\boldsymbol{\eta}_0^\top\boldsymbol{\eta}_1 + \bar{\mathbf{v}}_1^\top\boldsymbol{\eta}_1\,\bar{\mathbf{v}}_1^\top\boldsymbol{\eta}_0\Big)\\
&\qquad + r^{-\frac{5}{2}}\Big(\frac{(zm_2)'''}{3!}(zm_2) + \frac{(zm_2)''}{2}(zm_2)'\Big)\Big(\|\boldsymbol{\eta}_1\|^2 + \big(\bar{\mathbf{v}}_1^\top\boldsymbol{\eta}_1\big)^2\Big)\Big]\varphi_n(t) + O_{\prec}((|t|+1)n^{-\frac{1}{2}})\,.
\end{aligned} \tag{D.51}$$

Similarly, we also get

$$\begin{aligned}
\mathrm{i}\mathbb{E}\mathbb{T}_{32}h(t) &= -\frac{tm_1'}{(1+\sqrt{r}m_1)^2}\Big[r^{-\frac{3}{2}}(zm_2)'(zm_2)\Big(\boldsymbol{\eta}_0^\top\boldsymbol{\eta}_1 + \bar{\mathbf{v}}_1^\top\boldsymbol{\eta}_1\,\bar{\mathbf{v}}_1^\top\boldsymbol{\eta}_0\Big)\\
&\qquad + r^{-2}\Big(\frac{(zm_2)''}{2}(zm_2) + (zm_2)'(zm_2)'\Big)\Big(\|\boldsymbol{\eta}_1\|^2 + \big(\bar{\mathbf{v}}_1^\top\boldsymbol{\eta}_1\big)^2\Big)\Big]\varphi_n(t) + O_{\prec}((|t|+1)n^{-\frac{1}{2}})\,.
\end{aligned} \tag{D.52}$$

and

$$\begin{aligned}
\mathrm{i}\mathbb{E}\mathbb{T}_{33}h(t) &= \frac{t}{(1+\sqrt{r}m_1)}\Big[r^{-\frac{3}{2}}(zm_2)'(zm_2)\Big(\|\boldsymbol{\eta}_0\|^2 + \big(\bar{\mathbf{v}}_1^\top\boldsymbol{\eta}_0\big)^2\Big)\\
&\qquad + r^{-2}\Big(\frac{(zm_2)''}{2}(zm_2) + (zm_2)'(zm_2)'\Big)\Big(\boldsymbol{\eta}_1^\top\boldsymbol{\eta}_0 + \bar{\mathbf{v}}_1^\top\boldsymbol{\eta}_0\,\bar{\mathbf{v}}_1^\top\boldsymbol{\eta}_1\Big)\Big]\varphi_n(t) + O_{\prec}((|t|+1)n^{-\frac{1}{2}})\,.
\end{aligned} \tag{D.53}$$

Combining (D.43)- (D.45), (D.48)- (D.50) and (D.51)- (D.53), together with the definition of $\mathbf{y}_{0,1}, \tilde{\mathbf{y}}_{0,1}, \boldsymbol{\eta}_{0,1}$ in (D.31), after elementary computations, we can then conclude that

$$\varphi'(t) = \mathrm{i}\mathbb{E}\sum_{i,j=1}^{3}\mathbb{T}_{ij}h(t) = -\big(\mathbf{c}^\top\mathcal{M}\mathbf{c}\big)t\varphi_n(t) + O_{\prec}((|t|+1)n^{-\frac{1}{2}})\,.$$

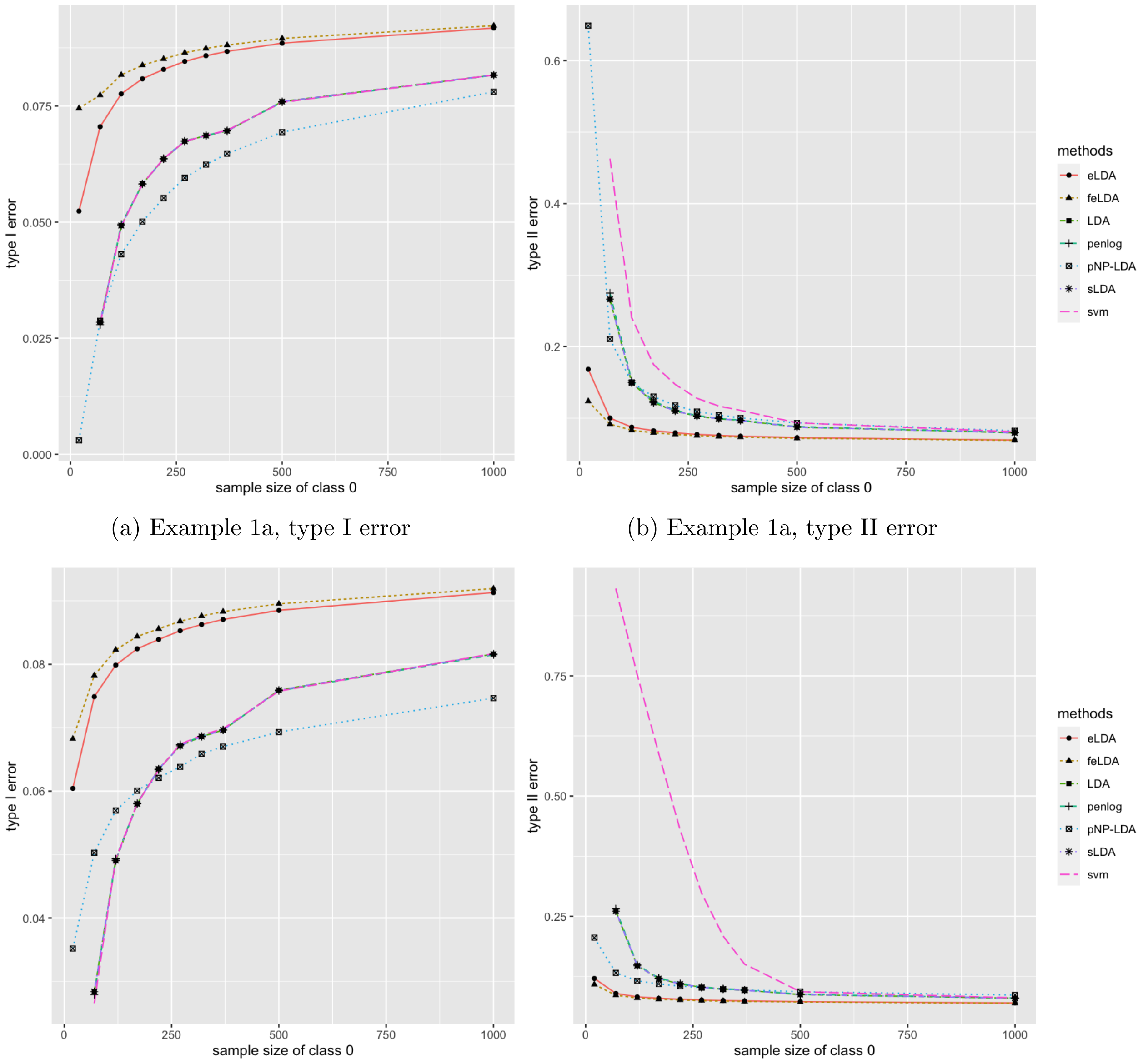
Hence, we finish the proof of Proposition D.1. ■

Appendix E. Additional numerical results

E.1 Additional figures and tables for simulation settings in Section 5

Figures 4, 5 and 6, Tables 2, 3 correspond to Examples 1 in Section 5. Figure 7 corresponds to Examples 2.

Figure 4: Examples 1a and 1b, type I and type II errors for competing methods with increasing balanced sample sizes (1a) and increasing n_0 only (1b) .



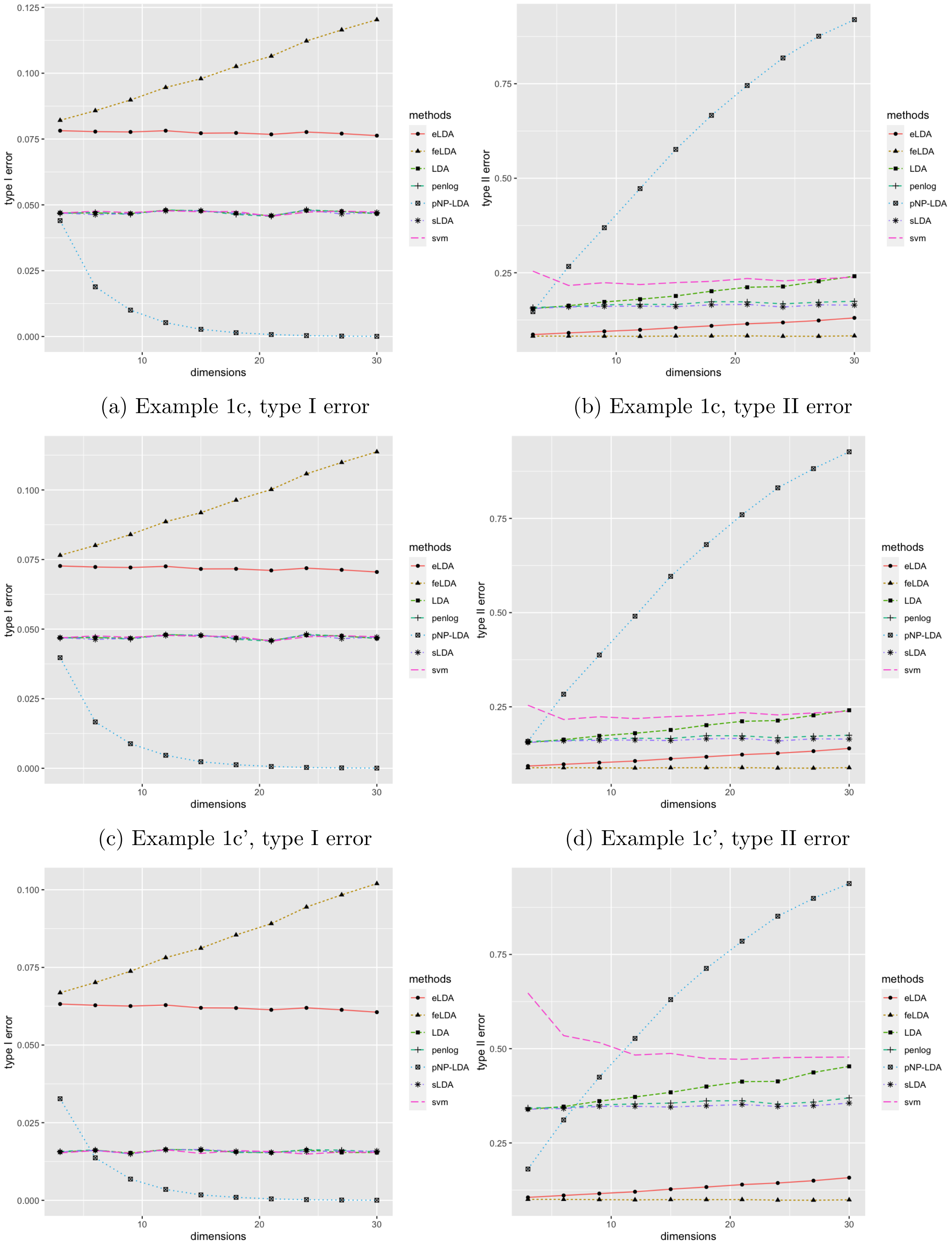
(a) Example 1a, type I error

(b) Example 1a, type II error

(c) Example 1b, type I error

(d) Example 1b, type II error

Figure 5: Examples 1c, 1c' and 1c*, type I and type II error for competing methods with increasing dimension p and different δ's: $\delta = 0.1$ in Example 1c, $\delta = 0.05$ in Example 1c', and $\delta = 0.01$ in Example 1c*.



(a) Example 1c, type I error

(b) Example 1c, type II error

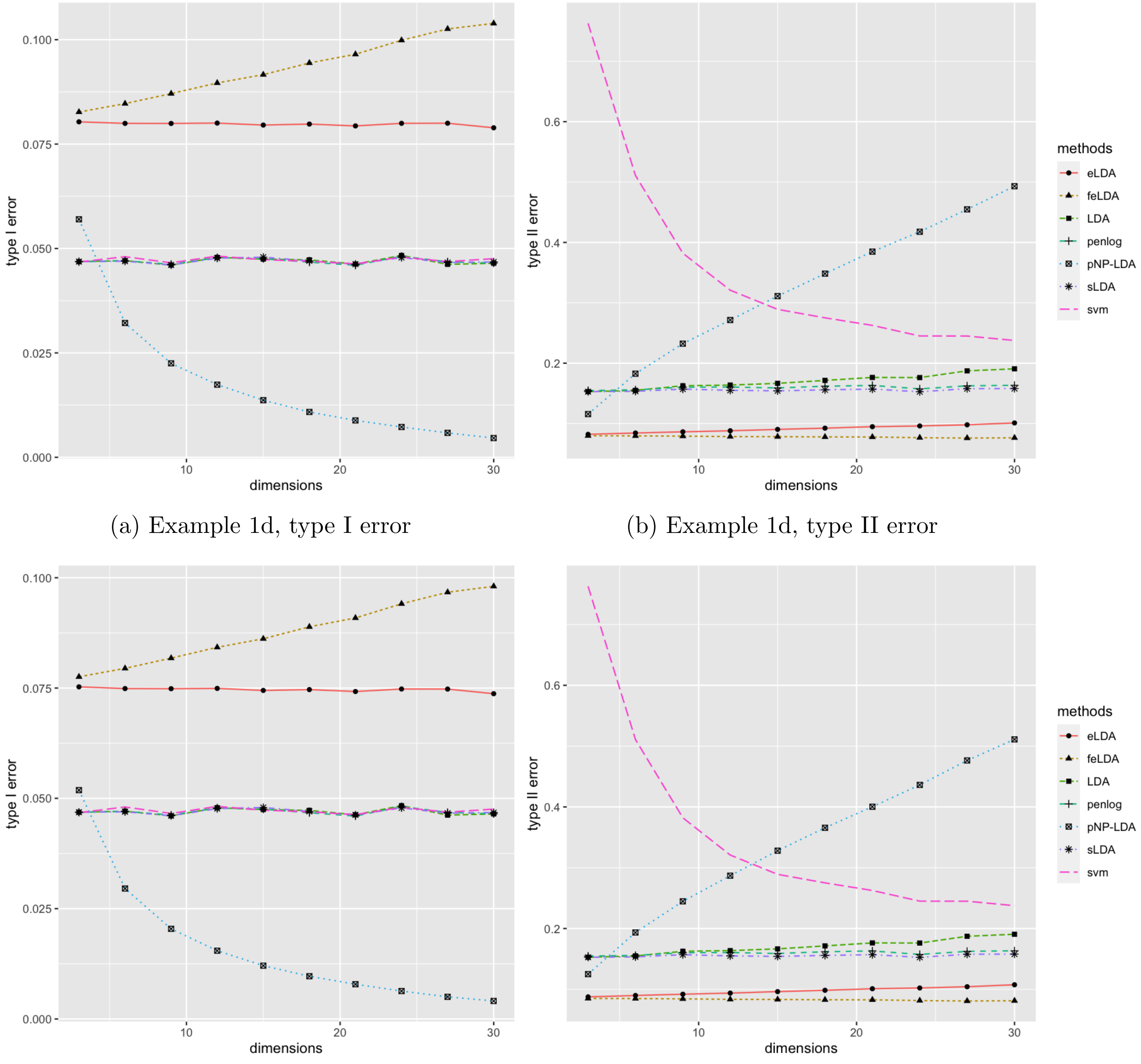
(c) Example 1c', type I error

(d) Example 1c', type II error

(e) Example 1c*, type I error

(f) Example 1c*, type II error

Figure 6: Examples 1d and 1d', imbalanced sample sizes with larger n_1. Type I and type II error for competing methods with increasing dimension p, but different δ's: $\delta = 0.1$ in Example 1d and $\delta = 0.05$ in Example 1d'.



(a) Example 1d, type I error

(b) Example 1d, type II error

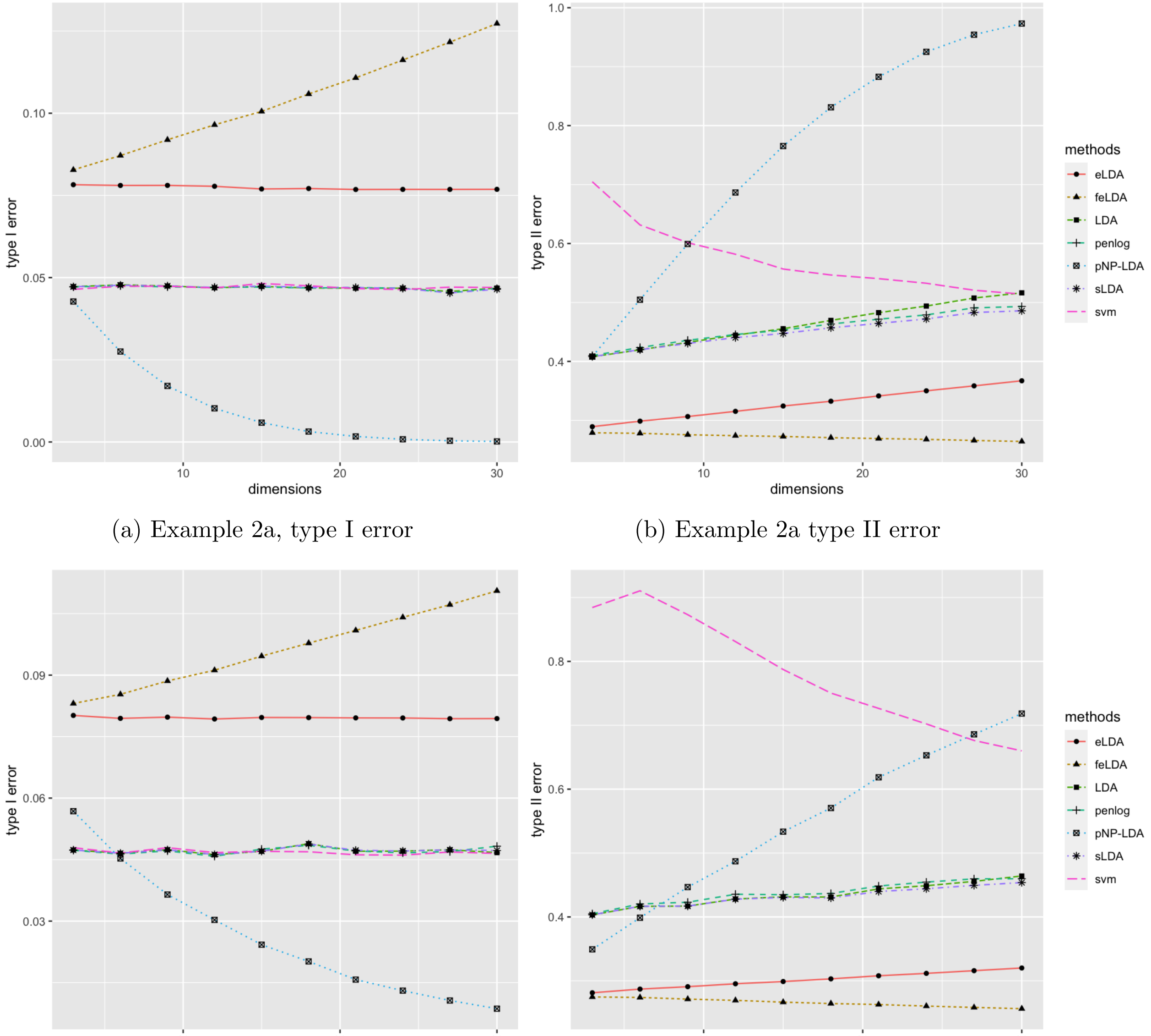
(c) Example 1d', type I error

(d) Example 1d', type II error

E.2 Lung cancer dataset continued

For the lung cancer dataset we explored in the real data section, we selected another set of parameters $\alpha = 0.1$, and $\delta = 0.4$ for a comparison among all five methods, including the umbrella algorithm based NP methods. We present the results in Table 10. We observe that, `eLDA` dominates `NP-slda`, `NP-penlog`, and `NP-svm` in both the type I and the type II

Figure 7: Examples 2a and 2b, type I and type II error for competing methods with increasing dimension p. Example 2a has balanced sample sizes and Example 2b has imbalanced sample sizes.



(a) Example 2a, type I error

(b) Example 2a type II error

(c) Example 2b, type I error

(d) Example 2b, type II error

errors. `pNP-lda` again produces a type I error of 0 and a type II error of 1: not informative at all. `eLDA` outperforms all other competing methods.

Table 10: Lung cancer dataset

		NP-slda	NP-penlog	NP-svm	pNP-lda	eLDA
$\alpha = 0.1$	type I error	.083	.078	.081	.000	.031
$\delta = 0.4$	type II error	.015	.026	.022	1.000	.013
	observed violation rate	.49	.45	.46	.00	.28

References

Alex Bloemendal, László Erdős, Antti Knowles, Horng-Tzer Yau, and Jun Yin. Isotropic local laws for sample covariance and generalized wigner matrices. *Electronic Journal of Probability*, 19:1–53, 2014.

Alex Bloemendal, Antti Knowles, Horng-Tzer Yau, and Jun Yin. On the principal components of sample covariance matrices. *Probability theory and related fields*, 164(1-2): 459–552, 2016.

Tony Cai and Linjun Zhang. High dimensional linear discriminant analysis: optimality, adaptive algorithm and missing data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(4):675–705, 2019.

Tony Cai and Linjun Zhang. A convex optimization approach to high-dimensional sparse quadratic discriminant analysis. *The Annals of Statistics*, 49(3):1537–1568, 2021.

Adam Cannon, James Howse, Don Hush, and Clint Scovel. Learning with the neyman-pearson and min-max criteria. *Los Alamos National Laboratory, Tech. Rep. LA-UR*, pages 02–2951, 2002.

László Erdős, Antti Knowles, and Horng-Tzer Yau. Averaging fluctuations in resolvents of random band matrices. *Ann. Henri Poincaré*, 14(8):1837–1926, 2013.

Jianqing Fan and Rui Song. Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38(6):3567–3604, 2010.

Jianqing Fan, Yang Feng, and Xin Tong. A road to classification in high dimensional space: the regularized optimal affine discriminant. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 74:745–771, 2012.

Jianqing Fan, Runze Li, Cun-hui Zhang, and Hui Zou. *Statistical Foundations of Data Science*. Chapman & Hall, 2020. ISBN 1466510846.

Yingying Fan, Yinfei Kong, Daoji Li, and Zemin Zheng. Innovated interaction screening for high-dimensional nonlinear classification. *The Annals of Statistics*, 43(3):1243–1272, 2015.

Gavin J Gordon, Roderick V Jensen, Li-Li Hsiao, Steven R Gullans, Joshua E Blumenstock, Sridhar Ramaswamy, William G Richards, David J Sugarbaker, and Raphael Bueno. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer research*, 62(17):4963–4967, 2002.

Ning Hao, Bin Dong, and Jianqing Fan. Sparsifying the fisher linear discriminant by rotation. *Journal of the Royal Statistical Society Series B*, 72:827–851, 2015.

Trevor Hastie, Robert Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd edition)*. Springer-Verlag Inc, 2009. ISBN 0-387-95284-5.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer New York, 2014. ISBN 9781461471370. URL `https://books.google.com.hk/books?id=at1bmAEACAAJ`.

Jiashun Jin and Wanjie Wang. Influential features pca for high dimensional clustering. *The Annals of Statistics*, 44(6):2323–2359, 2016.

Mohammadreza Kalan and Samory Kpotufe. Distribution-free rates in neyman-pearson classification. *https://arxiv.org/pdf/2402.09560.pdf*, 2024.

Quefeng Li and Jun Shao. Sparse quadratic discriminant analysis for high dimensional data. *Statistica Sinica*, pages 457–473, 2015.

Runze Li, Wei Zhong, and Liping Zhu. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139, 2012.

Yanfang Li and Jing Lei. Sparse subspace linear discriminant analysis. *Statistics*, 52(4): 782–800, 2018.

Qin Mai, Hui Zou, and Ming Yuan. A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, 99:29–42, 2012.

E. Mammen and A. Tsybakov. Smooth discrimination analysis. *Annals of Statistics*, 27: 1808–1829, 1999.

Manuel M Müller, Henry WJ Reeve, Timothy I Cannings, and Richard J Samworth. Isotonic subgroup selection. *arXiv preprint arXiv:2305.04852*, 2023.

Rui Pan, Hansheng Wang, and Runze Li. Ultrahigh-dimensional multiclass linear discriminant analysis by pairwise sure independence screening. *Journal of the American Statistical Association*, 111(513):169–179, 2016.

Yuqing Pan and Qing Mai. Efficient computation for differential network analysis with applications to quadratic discriminant analysis. *Computational Statistics & Data Analysis*, 144:106884, 2020.

W. Polonik. Measuring mass concentrations and estimating density contour clusters–an excess mass approach. *Annals of Statistics*, 23:855–881, 1995.

Henry WJ Reeve, Timothy I Cannings, and Richard J Samworth. Optimal subgroup selection. *The Annals of Statistics*, 51(6):2342–2365, 2023.

Philippe Rigollet and Xin Tong. Neyman-pearson classification, convexity and stochastic constraints. *Journal of Machine Learning Research*, 12(Oct):2831–2855, 2011.

Clayton Scott. Performance measures for neyman–pearson classification. *IEEE Transactions on Information Theory*, 53(8):2852–2863, 2007.

Clayton Scott. A generalized neyman-pearson criterion for optimal domain adaptation. *Proceedings of Machine Learning Research*, 93:1–24, 2019.

Clayton Scott and Robert Nowak. A neyman-pearson approach to statistical learning. *IEEE Transactions on Information Theory*, 51(11):3806–3819, 2005.

Jun Shao, Yazhen Wang, Xinwei Deng, and Sijian Wang. Sparse linear discriminant analysis by thresholding for high dimensional data. *Annals of Statistics*, 39:1241–1265, 2011.

Iman Sharafaldin, Arash Habibi Lashkari, and Ali A Ghorbani. Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp*, 1:108–116, 2018.

Houssem Sifaou, Abla Kammoun, and Mohamed-Slim Alouini. High-dimensional linear discriminant analysis classifier for spiked covariance model. *Journal of Machine Learning Research*, 21(112):1–24, 2020. URL `http://jmlr.org/papers/v21/19-428.html`.

Andrew I Su, John B Welsh, Lisa M Sapinoso, Suzanne G Kern, Petre Dimitrov, Hilmar Lapp, Peter G Schultz, Steven M Powell, Christopher A Moskaluk, Henry F Frierson, et al. Molecular classification of human carcinomas by use of gene expression signatures. *Cancer research*, 61(20):7388–7393, 2001.

Ye Tian and Yang Feng. Neyman-pearson multi-class classification via cost-sensitive learning. *arXiv:2111.04597*, 2021.

Xin Tong. A plug-in approach to neyman-pearson classification. *Journal of Machine Learning Research*, 14(1):3011–3040, 2013.

Xin Tong, Yang Feng, and Jingyi Jessica Li. Neyman-pearson classification algorithms and np receiver operating characteristics. *Science Advances*, 4(2):eaao1659, 2018.

Xin Tong, Lucy Xia, Jiacheng Wang, and Yang Feng. Neyman-pearson classification: parametrics and sample size requirement. *Journal of Machine Learning Research*, 21:1–18, 2020.

C. Wang and B. Jiang. On the dimension effect of regularized linear discriminant analysis. *Electronic Journal of Statistics*, 12:2709–2742, 2018.

Wanjie Wang, Jingjing Wu, and Zhigang Yao. Phase transitions for high-dimensional quadratic discriminant analysis with rare and weak signals. *arXiv preprint arXiv:2108.10802*, 2021.

Daniela Witten and Robert Tibshirani. Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society Series B*, 73:753–772, 2012.

Qing Yang and Guang Cheng. Quadratic discriminant analysis under moderate dimension. *arXiv:1808.10065*, 2018.

Mohammadmahdi R Yousefi, Jianping Hua, Chao Sima, and Edward R Dougherty. Reporting bias when using real data sets to analyze classification performance. *Bioinformatics*, 26(1):68–76, 2010.

Anqi Zhao, Yang Feng, Lie Wang, and Xin Tong. Neyman-pearson classification under high-dimensional settings. *Journal of Machine Learning Research*, 17(213):1–39, 2016.