Submitted to the Annals of Statistics

OPTIMAL CLASSIFICATION IN SPARSE GAUSSIAN GRAPHIC MODEL

By Yingying Fan*, Jiashun Jin†, and Zhigang Yao‡

University of Southern California, Carnegie Mellon University, and Ecole Polytechnique Fédérale de Lausanne

Consider a two-class classification problem when the number of features is much larger than the sample size. The features are masked by Gaussian noise with zero means and a covariance matrix Σ, where the precision matrix $\Omega = \Sigma^{-1}$ is unknown but is presumably sparse. The useful features, also unknown, are sparse and each contributes weakly (i.e., rare and weak) to the classification decision.

By obtaining a reasonably good estimate of Ω, we formulate the setting as a linear regression model. We propose a two-stage classification method where we first select features by the method of *Innovated Thresholding* (IT), and then use the retained features and Fisher's LDA for classification. In this approach, a crucial problem is how to set the threshold of IT. We approach this problem by adapting the recent innovation of Higher Criticism Thresholding (HCT).

We find that when useful features are rare and weak, the limiting behavior of HCT is essentially just as good as the limiting behavior of ideal threshold, the threshold one would choose if the underlying distribution of the signals is known (if only!). Somewhat surprisingly, when Ω is sufficiently sparse, its off-diagonal coordinates usually do not have a major influence over the classification decision.

Compared to recent work in the case where Ω is the identity matrix [15, 16], the current setting is much more general, which needs a new approach and much more sophisticated analysis. One key component of the analysis is the intimate relationship between HCT and Fisher's separation. Another key component is the tight large-deviation bounds for empirical processes associated with data with sparse but unconventional correlation structure, where the *separability of sparse graphs* plays an important role.

Keywords: Fisher's LDA, Fisher's separation, phase diagram, precision matrix, rare and weak model, separability of sparse graphs, sparse graph.

AMS 2000 subject classifications: Primary 62G05; secondary 62G32.

1. Introduction. Consider a two-class classification problem, where we have n labeled training samples $(X_i, Y_i), 1 \le i \le n$. Here, X_i are p-

*Supported in part by NSF CAREER Award DMS-1150318 and Grant DMS-0906784.
†Supported in part by NSF Grant DMS-1208315.
‡Supported in part by NSF Award SES-1061387 and NIH/NIDA Grant R90 DA023420.

dimensional feature vectors and $Y_i \in \{-1, 1\}$ are the corresponding class labels. For simplicity, we assume two classes are *equally likely*, and the data are centered so that

$$X_i \sim N(Y_i \cdot \mu,\ \Sigma_{p,p}), \tag{1.1}$$

where μ is the contrast mean vector between two classes, and $\Sigma_{p,p}$ is the $p \times p$ covariance matrix. Given a fresh feature vector

$$X \sim N(Y \cdot \mu,\ \Sigma_{p,p}), \tag{1.2}$$

the goal is to train (X_i, Y_i) to decide whether $Y = -1$ or $Y = 1$. We denote $\Sigma_{p,p}^{-1}$ by $\Omega_{p,p}$, and whenever there is no confusion, we drop the subscripts 'p, p' (and also that of any estimator of them, say, $\hat{\Omega}_{p,p}$).

We are primarily interested in the so-called '$p \gg n$' regime. In many applications where $p \gg n$ (e.g., genomics), we observe the following aspects.

- *Signals are rare.* Due to large p, the useful features (i.e., the nonzero coordinates of μ) are rare. For example, for a given type of cancer or disease, there are usually only a small number of relevant features (i.e., genes or proteins). When we measure increasingly more features, we tend to include increasingly more *irrelevant* ones.
- *Signals are individually weak.* The training data can be summarized by the z-vector

$$Z = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i X_i \sim N(\sqrt{n}\mu,\ \Sigma). \tag{1.3}$$

  Due to the small n, signals are weak in that, individually, the nonzero coordinates of $\sqrt{n}\mu$ are small or moderately large at most.
- *Precision matrix Ω is sparse.* Take Genetic Regulatory Network (GRN) for example. The feature vector $X = (X(1), \ldots, X(p))'$ represents the expression level of p different genes, and is approximately distributed as $N(\mu, \Sigma)$. For any $1 \le i \le p$, it is believed that for all except a few j, $1 \le j \le p$, the gene pair (i, j) are conditionally independent given all other genes. In other words, each row of Ω has only a few nonzero entries and so Ω is sparse [12].

In many applications, Ω is unknown and has to be estimated. In many other applications such as complicate disease or cancer, decades of biomedical studies have accumulated huge databases which are sometimes referred to as "data-for-data" [33]. Such databases can be used to accurately estimate Ω independently of the data at hand, and so Ω can be assumed as known.

In this paper, we investigate both the case where Ω is known and the case where Ω is unknown. In either case, we assume Ω has unit diagonals:

$$\Omega(i,i) = 1, \qquad 1 \le i \le p. \tag{1.4}$$

Such an assumption is only for simplicity, and we don't use such information for inference.

1.1. *Fisher's LDA and modern challenges.* Fisher's linear discriminant analysis (LDA) [20] is a well-known method for classification, which utilizes a weighted average of the test features $L(X) = \sum_{j=1}^{p} w(j)X(j)$, and predicts $Y = \pm 1$ if $L(X) >< 0$. Here, $w = (w(1), \ldots, w(p))'$ is a preselected weight vector. Fisher showed that the optimal weight vector satisfies

$$w \propto \Omega\mu. \tag{1.5}$$

In the classical setting where $n \gg p$, μ and Ω can be conveniently estimated and Fisher's LDA is approachable. Unfortunately, in the modern regime where $p \gg n$, Fisher's LDA faces immediate challenges.

- It is challenging to estimate Ω simply because that there are $O(p^2)$ unknown parameters but we have only $O(np)$ different measurements.
- Even when Ω is known and even in the simplest case where $\Omega = I_p$, challenges remain, as the signals are rare and weak. See [15] for the delicacy of the problem.

The paper is largely focused on addressing the second challenge, and shows successful classification can be achieved by simultaneously exploiting the sparsity of μ (aka. signal sparsity) and the sparsity of Ω (aka. graph sparsity). For the first challenge, encouraging progresses have been made recently (e.g., [21, 9]), and the problem is more or less settled. Still, the paper has a two-fold contribution along this line. First, we show that the performances of the methods in [21, 9] can be substantially improved if we add an additional re-fitting step; see details in Section 4. Second, we carefully analyze how the errors in estimating Ω may affect the classification results.

1.2. *Innovated Thresholding.* We wish to adapt Fisher's LDA to the current setting. Recall that the optimal choice of weight vector is $w \propto \Omega\mu$. If we have a reasonably good estimate of Ω (see Section 1.8 for more discussion on estimating Ω), say, $\hat{\Omega}$, all we need is a good estimate of μ.

When μ is sparse, one usually estimates it with some types of wavelet thresholding [41]. Let Z be the training z-vector as in (1.3). For some threshold t to be determined, there are three obvious approaches to thresholding:

- *Brute-force Thresholding (BT).* Applying thresholding to Z directly using the so-called clipping rule [15]: $\hat{\mu}_t^Z(i) = \mathrm{sgn}(Z(i))1\{|Z(i)| \geq t\}$ (alternatively, one may use soft-thresholding or hard thresholding [15], but the differences are secondary; similar below).
- *Whitened Thresholding (WT).* We first whiten the noise by the transformation $Z \mapsto \hat{\Omega}^{1/2}Z \approx N(\Omega^{1/2}\mu, I_p)$, and then apply the thresholding to the vector $\hat{\Omega}^{1/2}Z$ in a similar fashion.
- *Innovated Thresholding (IT).* We first take the transformation $Z \mapsto \hat{\Omega}Z$ and then apply the thresholding by

$$\hat{\mu}_t^{\hat{Z}}(i) = \mathrm{sgn}\big(\hat{Z}(i)\big)1\{|\hat{Z}(i)| \geq t)\}, \qquad \text{where } \hat{Z} \equiv \hat{\Omega}Z. \tag{1.6}$$

The transformation $Z \mapsto \hat{\Omega}Z$ is connected to the term of *Innovation* in the literature of time series [23], and so the name of Innovated Thresholding. Which of the three approaches is the best?

It turns out IT is the best. To see the point, note that for any $p \times p$ nonsingular matrix M, one could always estimate μ by applying the thresholding to MZ entry-wise (in BT, WT, and IT, $M = I_p, \Omega^{1/2}$, and Ω approximately). The deal is, what is the best M?

Towards this end, write $M = [m_1, m_2, \ldots, m_p]'$. For any $1 \leq i \leq p$, it is seen that $(MZ)(i) \sim N(\sqrt{n}m_i'\mu, m_i'\Sigma m_i)$. Therefore, if we bet on $\mu(i) \neq 0$, we should choose m_i to optimize the Signal to Noise Ratio (SNR) of $(MZ)(i)$. By Cauchy-Schwarz inequality, the optimal m_i satisfies that $m_i \propto \Omega\mu$. Writing $\Omega = [\omega_1, \omega_2, \ldots, \omega_p]$, it is seen that

$$\Omega\mu = \mu(i)\omega_i + \sum_{k \neq i} \mu(k)\omega_k \equiv (I) + (II). \tag{1.7}$$

When we bet on $\mu(i) \neq 0$, $(I) \propto \omega_i$ which is accessible to us. However, (II) is a very noisy vector and is inaccessible to us, estimating which is equally hard as estimating μ itself. The point can be further elaborated as follows: since we don't know the locations of other nonzero coordinates of μ, it makes sense to model $\{\sqrt{n}\mu(j) : 1 \leq j \leq p, j \neq i\}$ as *iid* samples from

$$(1 - \epsilon_p)\nu_0 + \epsilon_p H_p, \qquad \epsilon_p > 0: \quad \text{small}, \tag{1.8}$$

where ν_0 is the point mass at 0 and H_p is some distribution with no mass at 0. Under general "rare and weak" conditions for μ and sparsity condition for Ω, coordinates of $E[(II)]$ are uniformly small. This suggests that (II) is generally non-informative in designing the best m_i, and all we could utilize is (I).

In summary, if we bet on $\mu(i) \neq 0$, the optimal choice is $m_i \propto \omega_i$. As this holds for all i and we don't know where the signals are, the optimal choice for M is $M = \Omega$. This says that IT is not only the best among the three choices above, but is also the best choice in more general context.

In the literature of variable selection, IT is also called *marginal regression* [22]. The connection is not surprising, as approximately, $\hat{\Omega}^{1/2}Z \approx \Omega^{1/2}Z \sim N(\sqrt{n}\Omega^{1/2}\mu, I_p)$ which is a regression model. Both methods apply thresholding to ΩZ entry-wise, but marginal regression uses the hard thresholding rule, and IT uses the clipping thresholding rule [15].

With that being said, challenges remain on how to set the threshold t of IT (see (1.6)). If we set t too small or too large, the resultant estimator $\hat{\mu}_t^{\hat{Z}}$ has too many or too few nonzeros. Our proposal is to set the threshold in a data driven fashion by using the recent innovation of Higher Criticism Thresholding (HCT)

1.3. *Threshold choice by Higher Criticism.* Higher Criticism (HC) is a notion mentioned in passing by Tukey [40]. In recent years, HC was found to be useful in sparse signal detection [14], large-scale multiple testing [2, 7, 42], goodness-of-fit [29], and was applied to nonGaussian detection in Cosmic Microwave Background [11] and genomics [25, 35]. HC as a method for threshold choice in feature selection was first introduced in [15] (see also [24]), but the study has been focused on the case where Ω is the identity matrix. The case we consider in the current paper is much more complicated, where how to use HC for threshold choice is a non-trivial problem.

Our proposal is as follows. Let $\hat{\Omega}$ be a reasonably good estimate of Ω and let Z be the training z-vector as in (1.3). As in (1.6), denote for short

$$\hat{Z} = \hat{Z}(Z, \hat{\Omega}, p, n) = \hat{\Omega}Z. \tag{1.9}$$

The proposed approach contains three simple steps.

- For each $1 \leq j \leq p$, obtain a p-value by $\pi_j = P(|N(0,1)| \geq |\hat{Z}(j)|)$.
- Sort all the p-values in the ascending order $\pi_{(1)} < \pi_{(2)} < \ldots < \pi_{(p)}$.
- Define the HC functional $HC_{p,j} = \sqrt{p}[j/p - \pi_{(j)}]/\sqrt{(1-j/p)j/p}$, $1 \leq j \leq p$. Let $\hat{j}$ be the index at which $HC_{p,j}$ takes the maximum. The Higher Criticism Threshold (HCT)—denoted by $|\hat{Z}_{(\hat{j})}|$—is defined as the $\hat{j}$-th largest coordinate of $(|\hat{Z}(1)|, \ldots, |\hat{Z}(p)|)'$.

Moreover, for stability, we need the following refinement. Define

$$s_p^* = \sqrt{2\log(p)}, \qquad \tilde{s}_{p,n}^* = \tilde{s}_{p,n}^* = \sqrt{2\max\{0, \log(p/n^2)\}}. \tag{1.10}$$

It is well-understood (e.g., [14, 23]) that we should not allow the threshold to be larger than s_p^*. At the same time, we should not allow the threshold to be too small, especially when n is small. The Higher Criticism Threshold (HCT) we use in this paper is

$$t_p^{HC} = \begin{cases} |\hat{Z}_{(\hat{j})}|, & \text{if } \tilde{s}_{p,n}^* \le |\hat{Z}_{(\hat{j})}| \le s_p^*, \\ \tilde{s}_{p,n}^*, & \text{if } |\hat{Z}_{(\hat{j})}| < \tilde{s}_{p,n}^*, \\ s_p^*, & \text{if } |\hat{Z}_{(\hat{j})}| > s_p^*. \end{cases} \tag{1.11}$$

See Sections 1.5 and 3 for more detailed discussions.

1.4. *HCT trained classifier.* We are now ready for classification. Let $\hat{\Omega}$ be as above, and let $\hat{\mu}_{HC}^{\hat{Z}} = \hat{\mu}^{\hat{Z}}(Z, \hat{\Omega}, p, n)$ be defined as

$$\hat{\mu}_{HC}^{\hat{Z}}(j) = \mathrm{sgn}(\hat{Z}(j)) \cdot 1\{|\hat{Z}(j)| \ge t_p^{HC}\}, \qquad 1 \le j \le p. \tag{1.12}$$

Compared to $\hat{\mu}_t^{\hat{Z}}$ in (1.6), the only difference is that we have replaced t by t_p^{HC}. Introduce the HCT classification statistic

$$L_{HC}(X, \hat{\Omega}) = L_{HC}(X, \hat{\Omega}; Z, p, n) = (\hat{\mu}_{HC}^{\hat{Z}})' \hat{\Omega} Z. \tag{1.13}$$

The HCT trained classifier (or HCT classifier for short) is then the decision rule that decides $Y = \pm 1$ according to $L_{HC}(X, \hat{\Omega}) >< 0$.

The innovation of the procedure is two-fold: using IT for feature selection and using HCT for threshold choice in the more complicated case where Ω is unknown and is non-identity. The work is connected to other works on HC [23, 15], but the procedure and the delicate theory it entails are new.

A natural question is that whether IT has any advantages over exsiting variable selection methods (e.g., the Lasso [38], SCAD [19], Dantzig selector [10]). The answer is yes, for the following reasons. First, compared to these methods, IT is computationally much faster and much more approachable for delicate analysis. Second, our goal is classification, not variable selection. For classification, especially when features are rare and weak, the choice of different variable selection methods is secondary, while the choice of the tuning parameter is crucial. The threshold of IT can be conveniently set by HCT, but how to set the tuning parameter of the Lasso, SCAD, or Dantzig Selector remains an open problem, at least in theory.

How does the HCT classifier behave? In Sections 1.5-1.6, we set up a theoretic framework and derive a lower bound for classification errors. In Sections 1.7–1.8, we investigate the HCT classifier in the case where Ω is known and in the case where Ω is unknown separately, and show that the HCT classifier yields optimal phase diagram in classification.

1.5. *Asymptotic Rare and Weak model.* Motivated by the application examples aforementioned, we use a *Rare and Weak* signal model as follows. We model the scaled contrast mean vector $\sqrt{n}\mu$ as

$$\sqrt{n}\mu(j) \overset{iid}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p H_p, \qquad 1 \le j \le p, \tag{1.14}$$

where as in (1.8), ν_0 is the point mass at 0, H_p is some distribution with no mass at 0, and $\epsilon_p \in (0,1)$ is small (note that (ϵ_p, H_p) depend on p but not on j). We use p as the driving asymptotic parameter, and link (n, ϵ_p, H_p) to p through some fixed parameters. In detail, fixing parameters $(\beta, \theta) \in (0,1)^2$, we model

$$\epsilon_p = p^{-\beta}, \qquad n = n_p = p^{\theta}. \tag{1.15}$$

As p tends to ∞, the sample size n_p grows to ∞ but in a slower rate than that of p; the signals get increasingly sparser but the number of signals tends to ∞. The interesting range of parameters (β, θ, H_p) partitions into three regimes, according to the sparsity level.

- *Relatively Dense (RD).* In this regime, $0 < \beta < (1-\theta)/2$. The signals are relatively dense and successful classification is possible even when signals are very faint (e.g., H_p concentrates its mass around a term $\tau_p \ll \sqrt{2\log(p)}$). In such cases, (a) successful feature selection is impossible as signals are too weak, and (b) feature selection is unnecessary for the signals are relatively dense.
- *Rare and Weak (RW).* In this regime, $(1-\theta)/2 < \beta < (1-\theta)$, and the signals are moderately sparse. For successful classification, we need moderately strong signals (i.e., nonzero coordinates of $\sqrt{n}\mu \asymp \sqrt{\log(p)}$). In this case, feature selection is subtle but could be substantially helpful. In contrast, classification is impossible if signals are much weaker than $\sqrt{\log(p)}$, and feature selection is trivial if signals are much stronger than $\sqrt{\log(p)}$.
- *Rare and Strong (RS).* In this regime, $\beta > (1-\theta)$, and the signals are very sparse. For successful classification, we need very strong signals (signal strength $\gg \sqrt{\log(p)}$). In this case, feature selection is comparably easier to carry out (but substantially helpful) since the signals are strong enough to stand out for themselves.

While the statements hold broadly, the most transparent way to understand them is probably to consider the case where H_p is a point mass at τ_p (say): in the above three regimes, the minimum τ_p required for successful classification (up to some multi-$\log(p)$ factors in the first and last regimes)

are $1/(\epsilon_p\sqrt{(p/n_p)})$, $\sqrt{\log(p)}$, and $\sqrt{n_p/(p\epsilon_p)}$ correspondingly; the proof is elementary so is omitted.

In summary, feature selection is impossible in the RD regime and is relatively easy in the RS regime. For these reasons, we are primarily interested in the RW regime where we assume

$$(1-\theta)/2 < \beta < (1-\theta). \tag{1.16}$$

The RD/RS regimes are further discussed in Section 1.10, where we address the connection between our work and [18, 8]. For β in this range, the most interesting range for the signal strength is when H_p concentrates its mass at the scale of $\sqrt{\log(p)}$. In light of this, we fix $r > 0$ and calibrate the signal strength parameter τ_p by

$$\tau_p = \sqrt{2r\log(p)}. \tag{1.17}$$

Except in Section 1.6 where we address the lower bound arguments, we assume H_p is a point mass (compare (1.14)):

$$H_p = \nu_{\tau_p}, \qquad \text{where } \tau_p = \sqrt{2r\log(p)} \text{ is as in (1.17) and } 0 < r < 1. \tag{1.18}$$

We focus on the case $0 < r < 1$, as the case $r > 1$ corresponds to RS regime where the classification is comparably easier. This models a setting where the signal strengths are equal. The case where the signal strengths are unequal is discussed in Section 1.10.

Next, we model Ω. Motivated by the previous example on Genetic Regulatory Network, we assume each row of Ω has relatively few nonzeros. Such a matrix naturally induces a sparse graph $\mathcal{G} = (V, E)$, where $V = \{1, 2, \ldots, p\}$ and there is an edge between node i and j if and only if $\Omega(i, j) \neq 0$.

DEFINITION 1.1. *Fix $1 \leq K_p \leq p$. We call Ω K_p-sparse if and only if each row of Ω has at most K_p nonzeros, and we call $\mathcal{G}$ K_p-sparse if and only if the maximum degree $\leq K_p$.*

The class of K_p-sparse graphs is much broader than the class of banded graphs (we call $\mathcal{G}$ a banded graph with bandwidth K if nodes i and j are not connected whenever $|i-j| > K$). In fact, even when $\mathcal{G}$ is K_p-sparse with $K_p = 2$, we can not always shuffle the nodes of $\mathcal{G}$ and make it a banded graph with a small bandwidth.

Let $\mathcal{M}_p$ be the class of all $p \times p$ positive definite correlation matrices. Fixing $a \in (0, 1)$, $b > 0$, and a sequence of integers K_p, introduce

$$\mathcal{M}_p^*(a, K_p) = \{\Omega \in \mathcal{M}_p \text{ and is } K_p\text{-sparse}, \ |\Omega(i, j)| \leq a, \ i \neq j\}, \tag{1.19}$$

and

$$\widetilde{\mathcal{M}}_p^*(a,b,K_p) = \{\Omega \in \mathcal{M}_p^*(a,K_p), \|\Omega^{-1}\| \le b\}, \tag{1.20}$$

where $\|\cdot\|$ is the spectral norm. In comparison, $\widetilde{\mathcal{M}}_p^*(a,b,K_p)$ is slightly smaller than $\mathcal{M}_p^*(a,K_p)$. The following short-hand notation is frequently used in this paper.

DEFINITION 1.2. *We use L_p to denote a strictly positive generic multi-$\log(p)$ term that may vary from occurrence to occurrence but always satisfies that for any fixed $c > 0$, $\lim_{p\to\infty}\{L_p p^{-c}\} = 0$ and $\lim_{p\to\infty}\{L_p p^c\} = \infty$.*

In this paper, we are primarily interested in the case where K_p is at most multi-logarithmically large unless stated otherwise:

$$\lim_{p\to\infty} K_p = \infty, \qquad K_p \le L_p; \tag{1.21}$$

the first requirement is only for convenience. In our classification setting, $X_i \sim N(Y_i\mu, \Sigma)$, $X \sim N(Y\mu, \Sigma)$, and $Y = \pm 1$ with equal probabilities. The following notation is frequently used in the paper.

DEFINITION 1.3. *We say the classification problem (1.1)-(1.2) satisfies the Asymptotic Rare Weak model $ARW(\beta, r, \theta, \Omega)$ if (1.14)-(1.15), (1.18), and (1.21) hold.*

1.6. *Lower bound.* Introduce the the *standard phase boundary* function

$$\rho(\beta) = \begin{cases} 0, & 0 < \beta \le 1/2, \\ \beta - 1/2, & 1/2 < \beta < 3/4, \\ (1-\sqrt{1-\beta})^2, & 3/4 \le \beta < 1, \end{cases} \tag{1.22}$$

and let

$$\rho_\theta^*(\beta) = (1-\theta)\rho(\beta/(1-\theta)), \qquad (1-\theta)/2 < \beta < (1-\theta).$$

The function ρ has appeared before in determining phase boundaries in a seemingly unrelated problem on multiple hypothesis testing [26, 27, 14]. The following theorem is proved in Section 5.

THEOREM 1.1. *Fix $(\beta, r, \theta) \in (0,1)^3$ such that $(1-\theta)/2 < \beta < (1-\theta)$ and $0 < r < \rho_\theta^*(\beta)$. Suppose (1.14)-(1.15), (1.17), and (1.21) hold and that for sufficiently large p, $\Omega \in \mathcal{M}_p^*(a, K_p)$ and the support of H_p is contained in $[-\tau_p, \tau_p]$. Then as $p \to \infty$, for any sequence of trained classifiers, the misclassification error $\gtrsim 1/2$.*

Note that in Theorem 1.1, we don't require the signals to have the same strengths. Also, recall that in our classification setting (1.1)-(1.2), two classes are assumed as equally likely; extension to the case where two classes are unequally likely is straightforward. Theorem 1.1 was discovered before in [15, 28], but the study has been focused on the case where $\Omega = I_p$ and H_p is the point mass at τ_p. The proof in the current case is much more difficult and needs a few tricks, where *separability of sparse graphs* plays a key role.

LEMMA 1.1. *Fix a sufficiently large p and $1 \leq K_p < p$ and suppose $\mathcal{G} = (V, E)$ is a K_p-sparse graph. There is a constant $C > 0$ such that the graph decomposes into at most $CK_p \log(p)$ different disjoint subsets, where in each subset, there is no edge between any pair of nodes.*

Lemma 1.1 is proved in Section 5. The proof uses pigeon-hole principle and is elementary, but the result has far-reaching implications. Lemma 1.1 is the corner stone for proving the lower bound and for analyzing the HCT classifier (where we need tight convergence rate of empirical processes for data with non-conventional correlation structures).

1.7. *HCT achieves optimal phase diagram in classification (Ω is known).* One noteworthy aspect of HCT classifier is that it achieves the optimal phase diagram. In this section, we show this for the case where Ω is known. In this case, the HCT classifier $L_{HC}(X, \hat{\Omega})$ reduces to $L_{HC}(X, \Omega)$ (the term formed by replacing $\hat{\Omega}$ by Ω everywhere in the definition of former). The following theorem is proved in Section 5.

THEOREM 1.2. *Fix $(\beta, r, \theta, a) \in (0, 1)^4$ such that $(1-\theta)/2 < \beta < (1-\theta)$ and $r > \rho^*_\theta(\beta)$. Consider a sequence of classification problems $ARW(\beta, r, \theta, \Omega)$ with $\Omega \in \widetilde{\mathcal{M}}^*_p(a, b, K_p)$ for sufficiently large p. Then as p tends to ∞, $P\big(Y \cdot L_{HC}(X, \Omega) < 0\big) \to 0$. When $r < \beta$, the condition on Ω can be relaxed to that of $\Omega \in \mathcal{M}^*_p(a, K_p)$.*

Call the two-dimensional space $\{(\beta, r) : 0 < \beta < 1, 0 < r < 1\}$ the phase space. Theorems 1.1-1.2 say that the phase space partitions into two separate regions, *Region of Impossibility* and *Region of Possibility*, where the classification problem is distinctly different.

- *Region of Impossibility.* $\{(\beta, r) : (1-\theta)/2 < \beta < (1-\theta), 0 < r < \rho^*_\theta(\beta)\}$. Fix (β, r) in the interior of this region and consider a sequence of classification problems with $p^{1-\beta}$ signals where each signal $\leq \sqrt{2r\log(p)}$ in strength. Then for any sequence of 'sparse' Ω, suc-

cessful classification is impossible. This is the most difficult case where not much can be done for classification aside from random guessing.

- *Region of Possibility.* $\{(\beta, r) : (1-\theta)/2 < \beta < (1-\theta)\}, \rho^*_\theta(\beta) < r < 1\}$. Fix (β, r) in the interior of this region and suppose signals have equal strength of $\sqrt{2r\log(p)}$. HCT classifier $L_{HC}(X, \Omega)$ yields successful classification (the results hold much more broadly where equal signal strength assumption can be largely relaxed).

We call the curve $r = \rho^*_\theta(\beta)$ the *separating boundary*. Somewhat surprisingly, the separating boundary does not depend on the off-diagonals of Ω. The partition of phase diagram was discovered by [15, 31], and independently by [28], but where the focus was on the case where $\Omega = I_p$. See also [24]. The study in the current case is much more difficult. Similar phase diagram was also found in sparse signal detection [14], variable selection [30], and spectral clustering [32].

Why HCT works? The key insight is that there is an intimate relationship between the HC functional and Fisher's separation; the latter plays a key role in determining the optimal classification behavior, but is, unfortunately, an *oracle* quantity which depends on unknown parameters. In Sections 2–3, we outline a series of theoretic results, explaining why the HCT classifier is the right approach and how it achieves the optimality.

1.8. *Optimality of HCT classification (Ω is unknown).* When Ω is unknown, we first estimate it with the training data.

DEFINITION 1.4. *For any sequence of $\Omega_{p,p} \in \mathcal{M}^*_p(a, K_p)$, we say an estimator $\hat{\Omega}_{p,p}$ is acceptable if it is symmetric and independent of the test vector X, and that there is a constant $C > 0$ such that for sufficiently large p, $\hat{\Omega}_{p,p}$ is K'_p-sparse where $K'_p \le L_p$, and $|\hat{\Omega}_{p,p}(i,j) - \Omega_{p,p}(i,j)| \le CK_p^2\sqrt{\log(p)}/\sqrt{n_p}$ for all $1 \le i, j \le p$.*

Usually, the $(L_p/\sqrt{n_p})$-rate can not be improved, even when Ω is diagonal. For K_p-sparse Ω satisfying (1.21), acceptable estimators can be constructed based on the recent CLIME approach by [9]. If additionally Ω satisfies the mutual incoherence condition [34, Assumption 1], then the glasso [21] is also acceptable, provided the tuning parameters are properly set. If Ω is banded, then the Bickel and Levina Thresholding (BLT) method [4] is also acceptable, up to some modifications.

With that being said, the numeric performances of all these estimators can be improved with an additional step of *re-fitting*. See Section 4 for details.

Naturally, the estimation error of $\hat{\Omega}$ has some negative effects on the HCT classifier. Fortunately, for a large fraction of parameters (β, r) in Region of

Possibility, such effects are negligible and HCT continues to yield successful classification. In detail, suppose

- *Condition (a).* $r > \max\{(1-2\theta)/4, \rho^*_\theta(\beta)\}$,
- *Condition (b).* When $0 < \theta \le 1/3$ and $(1-\theta)/2 < \beta < (1-2\theta)$, $|r - \sqrt{1-2\theta}| \ge \sqrt{1-2\theta-\beta}$.

The following theorem is proved in Section 5.

THEOREM 1.3. *Fix $(\beta, r, \theta, a) \in (0,1)^4$ such that $(1-\theta)/2 < \beta < (1-\theta)$, and Conditions (a)-(b) hold. Consider a sequence of classification problems $ARW(\beta, r, \theta, \Omega)$ such that $\Omega \in \mathcal{M}^*_p(a, K_p)$ when $r < \beta$ and $\Omega \in \widetilde{\mathcal{M}}^*_p(a, b, K_p)$ when $r \ge \beta$. For the HCT classifier $L_{HC}(X, \hat{\Omega})$, if $\hat{\Omega}$ is acceptable, then as p tends to ∞, $P(Y \cdot L_{HC}(X, \hat{\Omega}) < 0) \to 0$.*

We remark that, first, when $0 < \theta \le 1/4$ and $(1-\theta)/2 < \beta < 3(1-2\theta)/4$, Condition (a) can be relaxed to that of $r > \max\{\beta/3, \rho^*_\theta(\beta)\}$. Second, when $\theta > 1/2$, Conditions (a)-(b) automatically hold when $r > \rho^*_\theta(\beta)$. As a result, we have the following corollary, the proof of which is omitted.

COROLLARY 1.1. *When $\theta > 1/2$, Theorem 1.3 holds with Conditions (a)-(b) replaced by that of $r > \rho^*_\theta(\beta)$.*

This says that as long as $n_p \gg \sqrt{p}$, the estimation errors of any acceptable estimator $\hat{\Omega}$ have negligible effects over the classification decision.

1.9. *Comparison with BT and WT.* In disguise, many methods are what we called 'Brute-forth Thresholding' or 'BT', including but not limited to [3, 17, 39]. Since Ω is hard to estimate, Bickel and Levina [3] and Tibshirani *et al* [39] neglect the off-diagonals in Σ for classification. In a seemingly different spirit, Efron [17] proposes a procedure where he first selects features by neglecting the off-diagonals in Σ and then estimates the correlation structures among selected features. However, under the Rare and Weak model, selected features tend to be uncorrelated. Therefore, at least for many cases, the approach fails to exploit the 'local' graphic structure of the data and is 'BT' in disguise. It is also noteworthy that [39] proposes to set the threshold of BT by cross validation, which is unstable, especially when n_p is small.

When we replace IT by either BT or WT in HCT classifier, the phase diagram associated with the resultant procedure is no longer optimal. While the claim holds very broadly, it can be conveniently illustrated with a simple case, where p is even, Ω is known and equals to the block diagonal matrix

calibrated by a parameter $h \in (-1,1)$ and where for all $1 \le i,j \le p$,

$$\Omega(i,j) = 1\{i=j\} + h\cdot 1\{j-i=1, i \text{ is odd}\} + h\cdot 1\{i-j=1, i \text{ is even}\}. \tag{1.23}$$

In this simple case, we have the following theorem, the proof of which is elementary so is omitted (a similar claim holds for WT if we replace $(1-h^2)$ by $(1+\sqrt{1-h^2})/2$ below).

THEOREM 1.4. *Fix $(\beta,\theta,r) \in (0,1)^3$ such that $(1-\theta)/2 < \beta < (1-\theta)$ and $\rho_\theta^*(\beta) < r < \rho_\theta^*(\beta)/(1-h^2)$. Suppose (1.18) and (1.23) hold. As $p \to \infty$, the classification error of HCT classifier tends to 0, but the HCT classifier with IT replaced by BT tends to 1/2, even when the threshold is ideally set.*

1.10. *Comparison with works focused on the RS Regime.* In disguise, many recent works focused on the "Rare and Strong" regime according to our terminology. One example is [36], where they assume the minimum signal strength (smallest coordinate in magnitude of $\sqrt{n_p}\mu$) is of the order of $\sqrt{n_p}$. Other examples include the ROAD approach by Fan *et al.* [18] and LPD approach by Cai *et al.* [8], where the main results (i.e., [18, Theorem 3], [8, Theorem 1]) assume a sparsity constraint that can be roughly translated to $\beta > (1-\theta/2)$ in our notations. Seemingly, this concerns the RS Regime we mentioned earlier.

Compared to these works, our work focuses on the most challenging regime where the signals are Rare and Weak, and we need much more sophisticated methods for feature selection and for threshold choices.

1.11. *Comparison with other popular classifiers.* HCT classifier also has advantages over well-known classifiers such as the Support Vector Machine (SVM) [6], Random Forest [5], and Boosting [13]. These methods need tuning parameters and are internally very complicated, but they do not outperform HCT classifier even when we replace the IT by BT; see details in [15], where we compared all these methods with three well-known gene microarray data sets in the context of cancer classification.

HCT is also closely related to PAM [39], but is different in important ways. First, HCT exploits the correlation structure while PAM does not. Second, while both methods perform feature selection, PAM sets the threshold by cross validations (CVT), while HCT sets the threshold by Higher Criticism. When n is small, CVT is usually unstable. In [15], we have shown that HCT outperforms CVT when analyzing the three microarray data sets aforementioned. In Section 4, we further compare HCT with CVT with simulated data.

1.12. *Summary and possible extensions.* We propose HCT classifier for two-class classification, where the major methodological innovation is the use of IT for feature selection and the use of HC for threshold choice.

IT is based on an 'optimal' linear transform that maximizes SNR in all signal locations, and has advantages over BT and WT. IT also has a three-fold advantages over the well-known variable selection methods such as the Lasso, SCAD, and Dantzig selector: (a) IT is computationally faster, (b) IT is more approachable in terms of delicate analysis, and (c) the tuning parameter of IT can be conveniently set, but how to set the tuning parameters of the other methods remains an open problem.

The idea of using HC for threshold choice goes back to [15], where the focus is on the case where Ω is known and is the identity matrix (see also [24]). In this paper, with considerable efforts, we extend the idea to the case where Ω is unknown but is presumably sparse, and show that HC achieves the optimal phase diagram in classification. The optimality of HC is not coincidental, and the underlying reason is the intimate relationship between the HC functional and Fisher's separation. This is explained in Section 2-3 with details.

In Theorems 1.2-1.3 and Section 2-3, we assume the signals have the same signs and strengths. The first assumption is largely for simplicity and can be removed. The second assumption can be largely relaxed, and both Theorems 1.2-1.3 and the intimate relationship between HC and Fisher's separation continue to hold to some extent if the signal strengths are unequal. One such example is where the signal distribution H_p, after scaled by a factor of $(\log(p))^{-1/2}$, has a continuous density over a closed interval contained in $(0, \infty)$ which does not depend on p.

In the paper, we also assume Ω (equivalently, the induced graph $\mathcal{G} = (V, E)$) is K-sparse for a moderately large K, which can also be relaxed. First, the main results continue to hold if there is an integer $M = M_p$ such that (a) $M_p \leq L_p$, and (b) V partitions into M different subsets, and any pair of nodes in the same subset are not connected (but nodes in different subsets could be connected in an arbitrary way). Second, when Ω have many small nonzero coordinates, we can always regularize it first with a threshold $t > 0$: $\Omega^*(i,j) = \Omega(i,j)1\{|\Omega(i,j)| \geq t\}$, and the main results continue to hold if Ω^* is K-sparse and the difference between two matrices is 'sufficiently small'.

1.13. *Content.* The remaining part of the paper is organized as follows. In Section 2, we introduce two functionals: Fisher's separation and ideal HC, and show that the two functionals are intimately connected to each other. In Section 3, we derive a large-deviation bound on the empirical cdf, and then

use it to characterize the stochastic fluctuation of the HC functional and that of Fisher's separation. Theorems 1.2-1.3 are proved in the end of this section. Section 4 contains numeric examples. Section 5 is the proof section, with proofs for secondary lemmas left to the appendix.

1.14. *Notations.* In this paper, $C > 0$ and $L_p > 0$ denote a generic constant and a generic multi-log(p) term respectively, which may vary from occurrence to occurrence. For two positive sequences $\{a_p\}_{p=1}^{\infty}$ and $\{b_p\}_{p=1}^{\infty}$, we say $a_p \sim b_p$ if $\lim_{p\to\infty}\{a_p/b_p\} = 1$ and we say $a_p \asymp b_p$ if there is a constant $c_0 > 1$ such that for sufficiently large p, $c_0^{-1} \le a_p/b_p \le c_0$.

The notations Ω and Σ are always associated with each other by $\Omega = \Sigma^{-1}$, and (X_i, Y_i) represents a training sample while (X, Y) represents a test sample. The summarizing z-vector for the training data set is denoted by Z, with $\tilde{Z} = \Omega Z$ and $\hat{Z} = \hat{\Omega} Z$, where $\hat{\Omega}$ is some estimate of Ω.

2. Ideal threshold and ideal HCT. In Sections 2-3, we discuss the behavior of HCT classifier. We limit our discussion to the $ARW(\beta, r, \theta, \Omega)$ model, but the key ideas are valid beyond the ARW model and extensions are possible; see discussions in Section 1.10.

The key insight behind the HCT methodology is that in a broad context,

$$\text{HCT} \approx \text{ideal HCT} \approx \text{ideal threshold}.$$

The ideal HCT is the non-stochastic counterpart of HCT, and the ideal threshold is the threshold one would choose if the underlying signal structure were known.

In this section, we elaborate the intimate connection between the ideal HCT and the ideal threshold, and their connections to Fisher's separation. We also investigate the performance of 'ideal classifier' where we assume Ω is known and the threshold is set ideally.

The connection between HCT and ideal HCT is addressed in Section 3, which is new even in the case of $\Omega = I_p$; compare [16]. Theorems 1.2-1.3 are also proved in Section 3.

2.1. *Fisher's separation and classification heuristics.* Fix a threshold $t > 0$ and let $\hat{\Omega}$ be an acceptable estimator of Ω. We are interested in the classifier that estimates $Y = \pm 1$ according to $L_t(X, \hat{\Omega}) >< 0$, where as in (1.12)-(1.13),

$$L_t(X, \hat{\Omega}) = (\hat{\mu}_t^{\hat{Z}})' \hat{\Omega} X \ \textit{with}\ \hat{\mu}_t^{\hat{Z}}(j) = \mathrm{sgn}(\hat{Z}(j)) 1\{|\hat{Z}(j)| \ge t\}.$$

For any fixed $p\times 1$ vector Z and $p\times p$ positive definite matrix A, we introduce

$$M_p(t,Z,\mu,A)=M_p(t,Z,\mu,A;n_p)=(\hat{\mu}_t^Z)'A\mu,$$

and

$$V_p(t,Z,A)=V_p(t,Z,A;\Omega)=(\hat{\mu}_t^Z)'A\Omega^{-1}A\hat{\mu}_t^Z,$$

where loosely, "M" and "V" stand for the mean and variance, respectively. In our model, given $(\mu,\hat{Z},\hat{\Omega})$, the test sample $X\sim N(Y\cdot\mu,\Omega^{-1})$; see (1.2) and note that $\hat{\Omega}$ is independent of X since it is acceptable. It follows that

$$L_t(X,\hat{\Omega})\sim N\big(Y\cdot M_p(t,\hat{Z},\mu,\hat{\Omega}),\ V_p(t,\hat{Z},\hat{\Omega})\big),$$

and the misclassification error rate of $L_t(X,\hat{\Omega})$ is

$$P(Y\cdot L_t(X,\hat{\Omega})<0|\mu,\hat{Z},\hat{\Omega})=\bar{\Phi}\bigg(\frac{M_p(t,\hat{Z},\mu,\hat{\Omega})}{\sqrt{V_p(t,\hat{Z},\hat{\Omega})}}\bigg), \tag{2.1}$$

where $\bar{\Phi}=1-\Phi$ denotes the survival function of $N(0,1)$.

The right hand side of (2.1) is closely related to the well-known Fisher's separation (Sep) [1], which measures the standardized interclass distance $Sep(t,\hat{Z},\mu,\hat{\Omega})=Sep(t,\hat{Z},\mu,\hat{\Omega};\Omega,p)$:

$$Sep(t,\hat{Z},\mu,\hat{\Omega};\Omega,p)=\frac{E[L_t(X,\hat{\Omega})|Y=1]-E[L_t(X,\hat{\Omega}))|Y=-1]}{SD(L_t(X,\hat{\Omega}))}. \tag{2.2}$$

In fact, it is seen that $Sep(t,\hat{Z},\mu,\hat{\Omega})=2M_p(t,\hat{Z},\mu,\hat{\Omega})/\sqrt{V_p(t,\hat{Z},\hat{\Omega})}$, and (2.1) can be rewritten as

$$P(Y\cdot L_t(X,\hat{\Omega})<0|\mu,\hat{Z},\hat{\Omega})=\bar{\Phi}\big(\frac{1}{2}Sep(t,\hat{Z},\mu,\hat{\Omega})\big).$$

By (1.14) and (1.18), the overall misclassification error rate is then

$$P(Y\cdot L_t(X,\hat{\Omega})<0)=E_{\epsilon_p,\tau_p}E\bigg[\bar{\Phi}\big(\frac{1}{2}Sep(t,\hat{Z},\mu,\hat{\Omega})\big)\bigg], \tag{2.3}$$

where E is the expectation with respect to the law of $(\hat{Z},\hat{\Omega}|\mu)$, and E_{ϵ_p,τ_p} is the expectation with respect to the law of μ; see (1.14) and (1.18).

We introduce two proxies for Fisher's separation. Throughout this paper,

$$\tilde{Z}=\Omega Z. \tag{2.4}$$

For the first proxy, recall that $\hat{Z} = \hat{\Omega}Z$ (e.g., (1.9)). Heuristically, $\hat{\Omega} \approx \Omega$ and so $\hat{Z} \approx \tilde{Z}$. We expect that $Sep(t, \hat{Z}, \mu, \hat{\Omega}) \approx Sep(t, \tilde{Z}, \mu, \Omega)$; the latter is Fisher's separation for the idealized case where Ω is known and is defined as

$$Sep(t, \tilde{Z}, \mu, \Omega) = 2M_p(t, \tilde{Z}, \mu, \Omega)/\sqrt{V_p(t, \tilde{Z}, \Omega)}. \tag{2.5}$$

For the second proxy, we note that when p is large, some regularity appears, and we expect that $M_p(t, \tilde{Z}, \mu, \Omega) \approx m_p(t, \epsilon_p, \tau_p, \Omega)$ and $V_p(t, \tilde{Z}, \Omega) \approx v_p(t, \epsilon_p, \tau_p, \Omega)$, where

$$m_p(t, \epsilon_p, \tau_p, \Omega) = E[M_p(t, \tilde{Z}, \mu, \Omega)], \qquad v_p(t, \epsilon_p, \tau_p, \Omega) = E[V_p(t, \tilde{Z}, \Omega)]. \tag{2.6}$$

In light of this, a second proxy separation is the *population Sep*:

$$\widetilde{Sep}(t) = \widetilde{Sep}(t, \epsilon_p, \tau_p, \Omega) = 2m_p(t, \epsilon_p, \tau_p, \Omega)/\sqrt{v_p(t, \epsilon_p, \tau_p, \Omega)}.$$

In summary, we expect to see that

$$Sep(t, \hat{Z}, \mu, \hat{\Omega}) \approx Sep(t, \tilde{Z}, \mu, \Omega) \approx \widetilde{Sep}(t, \epsilon_p, \tau_p, \Omega),$$

and that

$$P(Y \cdot L_t(X, \hat{\Omega}) < 0) \approx \bar{\Phi}\big(\frac{1}{2}\widetilde{Sep}(t)\big). \tag{2.7}$$

In Section 3, we solidify the above connections. But before we do that, we study the ideal threshold—the threshold that maximizes $\widetilde{Sep}(t)$.

2.2. *Ideal threshold.* Ideally, one would choose t to minimize the classification error of $L_t(X, \hat{\Omega})$. In light of (2.7), this is almost equivalent to choosing t as the ideal threshold.

DEFINITION 2.1. *The ideal threshold $T_{ideal}(\epsilon_p, \tau_p, \Omega)$ is the maximizing point of the second proxy: $T_{ideal}(\epsilon_p, \tau_p, \Omega) = \mathrm{argmax}_{\{0<t<\infty\}} \widetilde{Sep}(t, \epsilon_p, \tau_p, \Omega)$.*

In general, $\widetilde{Sep}(t, \epsilon_p, \tau_p, \Omega)$ and $T_{ideal}(\epsilon_p, \tau_p, \Omega)$ may depend on Ω in a complicated way. Fortunately, it turns out that for large p and all Ω in $\mathcal{M}_p^*(a, K_p)$ (see (1.19)), the leading terms of $\widetilde{Sep}(t)$ and $T_{ideal}(\epsilon_p, \tau_p, \Omega)$ do not depend on the off-diagonals of Ω and have rather simple forms.

DEFINITION 2.2. *(Folding). Denote $\Psi_\tau(t) = P(|N(\tau, 1)| \le t)$. When $\tau = 0$, we drop the subscript and write $\Psi(t)$. Also, denote $\bar{\Psi}_\tau = 1 - \Psi_\tau(t)$ and $\bar{\Psi}(t) = 1 - \Psi(t)$.*

In detail, let

$$\widetilde{W}_0(t) = \widetilde{W}_0(t, \epsilon_p, \tau_p; \Psi) = \epsilon_p \bar{\Psi}_{\tau_p}(t) / \sqrt{\bar{\Psi}(t) + \epsilon_p \bar{\Psi}_{\tau_p}(t)}, \tag{2.8}$$

$$t_p^*(\beta, r) = \min\{2, \frac{r+\beta}{2r}\}\tau_p, \tag{2.9}$$

and

$$\delta(\beta, r) = \begin{cases} \beta - r, & r \le \beta/3, \\ \frac{(\beta+r)^2}{8r}, & \beta/3 < r < \beta, \\ \beta/2, & \beta \le r < 1. \end{cases} \tag{2.10}$$

Elementary calculus shows that for large p,

$$\text{argmax}_{\{0\le t<\infty\}}\{\widetilde{W}_0(t)\} \sim t_p^*(\beta, r), \quad \sup_{\{0\le t<\infty\}} \widetilde{W}_0(t) = L_p \cdot p^{-\delta(\beta,r)}. \tag{2.11}$$

It turns out that there is an intimate relationship between $\widetilde{Sep}(t, \epsilon_p, \tau_p, \Omega)$ and $\widetilde{W}_0(t, \epsilon_p, \tau_p)$, where the latter does not depend on the off-diagonals of Ω. To see the point, we discuss the cases $r < \beta$ and $r \ge \beta$ separately.

In the first case, for a as in $\mathcal{M}_p^*(a, K_p)$, we let

$$c_0(\beta, r, a) = \delta(\beta, r) - \delta(\beta, a^2 r), \qquad \tilde{c}_0(\beta, r, a) = \tilde{c}_1(\beta, r, a) - \delta(\beta, r), \tag{2.12}$$

where if $a < 1/3$, $\tilde{c}_1(\beta, r, a) = \beta$, and otherwise,

$$\tilde{c}_1(\beta, r, a) = \begin{cases} \frac{(3a-1)r}{3-a} + \beta, & r \le \frac{3-a}{1+5a}\beta, \\ \frac{3-a}{1+a}\frac{(\beta+r)^2}{8r}, & \frac{3-a}{1+5a}\beta < r \le \beta. \end{cases}$$

The following lemma is proved Section 5.

LEMMA 2.1. *Fix $(\beta, r, \theta, a) \in (0,1)^4$ such that $\rho_\theta^*(\beta) < r < \beta$ and $(1-\theta)/2 < \beta < (1-\theta)$. In the ARW$(\beta, r, \theta, \Omega)$ model, as $p \to \infty$,*

$$\sup_{t>0} \sup_{\{\Omega \in \mathcal{M}_p^*(a, K_p)\}} \big|p^{\frac{(\theta-1)}{2}} \widetilde{Sep}(t, \epsilon_p, \tau_p, \Omega) - 2\tau_p \widetilde{W}_0(t, \epsilon_p, \tau_p)\big| \le L_p p^{-\max\{\beta - \frac{r}{2}, \frac{3\beta+r}{4}\}}$$
$$+ L_p\big[p^{-\min\{r, \frac{\beta-r}{2}, (1-a)(\beta-ar)\}} + p^{-c_0(\beta,r,a)} + p^{-\tilde{c}_1(\beta,r,a)}\big] \sup_{\{0<t<\infty\}} \widetilde{W}_0(t, \epsilon_p, \tau_p).$$

Compared to the left hand side, the right hand side is much smaller and is negligible. Therefore, approximately, $\widetilde{Sep}(t,\epsilon_p,\tau_p,\Omega) \propto \widetilde{W}_0(t,\epsilon_p,\tau_p)$ for all $\Omega \in \mathcal{M}_p^*(a,K_p)$. Combining this with (2.11), we expect to have

$$T_{ideal}(\epsilon_p,\tau_p,\Omega) \sim t_p^*(\beta,r), \quad \sup_{0<t<\infty} \widetilde{Sep}(t,\epsilon_p,\tau_p,\Omega) = L_p p^{\frac{1-\theta}{2}-\delta(\beta,r)}. \tag{2.13}$$

Next, consider the case $r \geq \beta$. The lemma below is proved in Section 5.

LEMMA 2.2. *Fix $(\beta,r,\theta,a) \in (0,1)^4$ such that $r \geq \beta$ and $(1-\theta)/2 < \beta < (1-\theta)$. Let $\Delta_1 = d_0 \log(\log(p))/\sqrt{\log p}$ and $\Delta_2 = 2\sqrt{\log(K_p \log p)}$, where $d_0 > 0$ is some constant. In the ARW(β,r,θ,Ω) model with $\Omega \in \widetilde{\mathcal{M}}_p^*(a,b,K_p)$, as $p \to \infty$,*

(a) $\sup_{\{0<t<\sqrt{2\beta\log(p)}-\Delta_1\}} \widetilde{Sep}(t,\epsilon_p,\tau_p,\Omega) \lesssim \frac{5}{3}\tau_p K_p^{-1} p^{(1-\theta-\beta)/2}$,

(b) $\sup_{\{t\geq\tau_p+\Delta_2\}} \widetilde{Sep}(t,\epsilon_p,\tau_p,\Omega) \lesssim \frac{5}{3}\tau_p K_p^{-1} p^{(1-\theta-\beta)/2}$,

(c) $2\tau_p K_p^{-1} p^{\frac{1-\theta-\beta}{2}} \lesssim \sup_{\{\sqrt{2\beta\log p}-\Delta_1\leq t<\tau_p\}} \widetilde{Sep}(t,\epsilon_p,\tau_p,\Omega) \leq L_p p^{\frac{1-\theta-\beta}{2}}$.

A direct result of Lemma 2.2 is that, for all $\Omega \in \widetilde{\mathcal{M}}_p^*(a,b,K_p)$ (see (1.19)),

$$\sqrt{2\beta\log(p)} \lesssim T_{ideal} \lesssim \sqrt{2r\log(p)}, \quad \sup_{\{0<t<\infty\}} \{\widetilde{Sep}(t)\} \asymp L_p p^{(1-\theta-\beta)/2}, \tag{2.14}$$

where $T_{ideal} = T_{ideal}(\epsilon_p,\tau_p,\Omega)$ and $\widetilde{Sep}(t) = \widetilde{Sep}(t,\epsilon_p,\tau_p,\Omega)$ for short. In this case, the function $\widetilde{Sep}(t)$ sharply increases and decreases in the intervals $(0,\sqrt{2\beta\log(p)})$ and $(\sqrt{2r\log(p)},\infty)$, respectively, but is relatively flat in the interval $(\sqrt{2\beta\log(p)},\sqrt{2r\log(p)})$; in this interval, the function reaches the maximum but varies slowly at the magnitude of $O(L_p p^{(1-\theta-\beta)/2})$. In the current case, on one hand, it is not critical to pin down T_{ideal}, as $\widetilde{Sep}(t) = L_p p^{(1-\theta-\beta)/2}$ for all t in the whole interval. On the other hand, it is hard to pin down T_{ideal} uniformly for all Ω under consideration, if possible at all.

2.3. *Ideal HCT.* Ideal HCT is a counterpart of HCT and a non-stochastic threshold that HCT tries to estimate. Introduce a functional which is defined over all survival functions associated with a positive random variable:

$$HC(t,G) = \sqrt{p}[G(t) - \bar{\Psi}(t)]/\sqrt{G(t)(1-G(t))}, \qquad t > 0.$$

We are primarily interested in thresholds that are neither too small or too large as far as HCT concerns; see (1.10). In light of this, we introduce the HCT functional

$$T_{HC}(G) = \mathrm{argmax}_{\{\bar{\Psi}^{-1}(\frac{1}{2})<t<s_p^*\}} HC(t,G),$$

where the term $\bar{\Psi}^{-1}(1/2)$ is chosen for convenience, and can be replaced by some other positive constants. Recall that $\tilde{Z} = \Omega Z$ and $\hat{Z} = \hat{\Omega} Z$ (e.g., (2.4) and 1.9)). For any $t > 0$, let

$$\bar{F}_p(t) = \frac{1}{p} \sum_{j=1}^{p} 1\{|\hat{Z}(j)| \geq t\}, \tag{2.15}$$

and

$$\widetilde{F}_p(t) = \frac{1}{p} \sum_{j=1}^{p} 1\{|\tilde{Z}(j)| \geq t\}, \quad \widetilde{F}(t) = \widetilde{F}(t, \epsilon_p, \pi_p, \Omega) = E_{\epsilon_p, \pi_p}[\widetilde{F}_p(t)]. \tag{2.16}$$

Note that the only difference between $\widetilde{F}_p(t)$ and $\widetilde{F}(t)$ is the subscript p. Heuristically, for large p, we expect to have $\bar{F}_p(t) \approx \widetilde{F}_p(t) \approx \widetilde{F}(t)$. As a result, we expect that

$$T_{HC}(\bar{F}_p) \approx T_{HC}(\widetilde{F}_p) \approx T_{HC}(\widetilde{F}),$$

where $T_{HC}(\bar{F}_p)$ is the HCT where Ω is unknown and has to be estimated, $T_{HC}(\widetilde{F}_p)$ is the HCT when Ω is known, and $T_{HC}(\widetilde{F})$ is a non-stochastic counterpart of $T_{HC}(\widetilde{F}_p)$.

DEFINITION 2.3. *We call $T_{HC}(\widetilde{F})$ the ideal Higher Criticism Threshold (ideal HCT).*

Similarly, the leading term of ideal HCT has a simple form that is easy to analyze. Fix $1 \leq j \leq p$. Let $D_j = \{k : 1 \leq k \leq p, \Omega(j,k) \neq 0\}$, and let

$$g_1(t) = g_1(t; \Omega, \epsilon_p, \tau_p) = \frac{1}{p} \sum_{j=1}^{p} P\big(|\tilde{Z}(j)| \geq t, \mu(k) \neq 0 \text{ for some } k \in D_j,\ k \neq j\big).$$

The following is a counterpart of $\widetilde{W}_0(t)$ defined in (2.8) and can be well approximated by the latter:

$$W_0(t) = W_0(t, \epsilon_p, \tau_p, \Omega) = \frac{\epsilon_p \bar{\Psi}_{\tau_p}(t) + g_1(t)}{\sqrt{\bar{\Psi}(t) + \epsilon_p \bar{\Psi}_{\tau_p}(t) + g_1(t)}}. \tag{2.17}$$

The following lemmas are proved in Section 5.

LEMMA 2.3. *Fix $(\beta, r, \theta, a) \in (0,1)^4$ such that $r > \rho^*_\theta(\beta)$ and $(1-\theta)/2 < \beta < (1-\theta)$. In the ARW$(\beta, r, \theta, \Omega)$ model, as $p \to \infty$,*

$$\sup_{\{t > \bar{\Psi}^{-1}(\frac{1}{2})\}} \sup_{\{\Omega \in \mathcal{M}^*_p(a, K_p)\}} \{\big|p^{-1/2} HC(t, \widetilde{F}) - W_0(t, \epsilon_p, \tau_p, \Omega)\big|\} \leq L_p p^{-\beta}.$$

LEMMA 2.4. *Fix $(\beta, r, \theta, a) \in (0,1)^4$ such that $r > \rho^*_\theta(\beta)$ and $(1-\theta)/2 < \beta < (1-\theta)$. In the ARW$(\beta, r, \theta, \Omega)$ model, as $p \to \infty$, we have*

$$\sup_{\{t>0\}} \sup_{\{\Omega \in \mathcal{M}^*_p(a, K_p)\}} \big|W_0(t, \epsilon_p, \tau_p, \Omega) - \widetilde{W}_0(t, \epsilon_p, \tau_p)\big| \le L_p\big[p^{-\frac{3\beta}{2}} + p^{-c_0(\beta,r,a)} \sup_{\{t>0\}} \widetilde{W}_0(t)\big].$$

If additionally $r \ge \beta$, then

(a) $\sup_{\{0 \le t < \sqrt{2\beta\log(p)} - \Delta_1\}} W_0(t, \epsilon_p, \tau_p, \Omega) \lesssim (\frac{1}{\sqrt{2}}) p^{-\beta/2}$,
(b) $\sup_{\{\tau_p \le t < \infty\}} W_0(t, \epsilon_p, \tau_p, \Omega) \lesssim (\frac{1}{\sqrt{2}}) p^{-\beta/2}$,
(c) $p^{-\beta/2} \lesssim \sup_{\{\sqrt{2\beta\log(p)} - \Delta_1 < t < \tau_p\}} W_0(t, \epsilon_p, \tau_p, \Omega) \le L_p p^{-\beta/2}$,

where $\Delta_1 = d_0 \log\log(p)/\sqrt{\log(p)}$ is defined in Lemma 2.2.

Lemmas 2.3-2.4 say that, approximately, $HC(t, \widetilde{F}) \propto W_0(t)$, and that two functions $\widetilde{W}_0(t)$ and $W_0(t)$ are generally close.

Together, Lemmas 2.1-2.4 consolidate the intimate relationship between the ideal threshold and the ideal HCT. To see the point, we discuss the cases $r < \beta$ and $r \ge \beta$ separately.

For the first case, write $T_{ideal} = T_{ideal}(\epsilon_p, \tau_p, \Omega)$ and $\widetilde{Sep}(t) = \widetilde{Sep}(t, \epsilon_p, \tau_p, \Omega)$ for short as before. The following theorem is proved in Section 5.

THEOREM 2.1. *Fix $(\beta, r, \theta, a) \in (0,1)^4$ such that $\rho^*_\theta(\beta) < r < \beta$ and $(1-\theta)/2 < \beta < (1-\theta)$. In the ARW$(\beta, r, \theta, \Omega)$ model with $\Omega \in \mathcal{M}^*_p(a, K_p)$, as $p \to \infty$, there is a constant $c_1 = c_1(\beta, r, a) > 0$ such that $|T_{HC}(\widetilde{F}) - T_{ideal}| \le L_p p^{-c_1(\beta,r,a)}$, and so $\widetilde{Sep}(T_{HC}(\widetilde{F})) \sim \widetilde{Sep}(T_{ideal}) = L_p p^{(1-\theta)/2 - \delta(\beta,r)}$.*

Consider the second case. Lemmas 2.4 says that $\sqrt{2\beta\log(p)} \lesssim T_{HC}(\widetilde{F}) \lesssim \sqrt{2r\log(p)}$. While it is hard to further elaborate how close two ideal thresholds are, in light of (2.14), classification by ideal HCT is at least "sub-optimal". The following theorem is proved in Section 5.

THEOREM 2.2. *Fix $(\beta, r, \theta, a) \in (0,1)^4$ such that $r \ge \beta$ and $(1-\theta)/2 < \beta < (1-\theta)$. In the ARW$(\beta, r, \theta, a)$ model where $\Omega \in \widetilde{\mathcal{M}}^*_p(a, b, K_p)$, as $p \to \infty$, we have that $2\tau_p K_p^{-1} p^{(1-\theta-\beta)/2} \lesssim \widetilde{Sep}(T_{HC}(\widetilde{F})) \le \widetilde{Sep}(T_{ideal}(\epsilon_p, \tau_p, \Omega)) = L_p p^{(1-\theta-\beta)/2}$.*

To conclude this section, we investigate the 'ideal' classifier $L_t(X, \Omega)$, where Ω is known to us. Note that for each fixed t, the misclassification error of $L_t(X, \Omega)$ is $P(Y \cdot L_t(X, \Omega) < 0) = E_{\epsilon_p, \pi_p} E\big[\bar{\Phi}(\frac{1}{2} Sep(t, \tilde{Z}, \mu, \Omega)\big]$. The following theorem is proved in Section 5.

THEOREM 2.3. *Fix $(\beta, r, \theta, a) \in (0,1)^4$ such that $(1-\theta)/2 < \beta < (1-\theta)$ and $r > \rho^*_\theta(\beta)$. In the ARW(β, r, θ, a) model with $\Omega \in \widetilde{\mathcal{M}}^*_p(a, b, K_p)$, as $p \to \infty$,*

$$\min_t P\big(Y \cdot L_t(X,\Omega) < 0|t\big) = \bar{\Phi}\bigg((1+o(1)) \cdot \frac{1}{2}\widetilde{Sep}\big(T_{ideal}\big)\bigg).$$

When $r < \beta$, the condition $\Omega \in \widetilde{\mathcal{M}}^*_p(a, b, K_p)$ can be relaxed to that of $\Omega \in \mathcal{M}^*_p(a, K_p)$.

Combining Theorem 2.3 with Theorems 2.1-2.2,

$$\min_t P\big(Y \cdot L_t(X,\Omega) < 0|t\big) = \bar{\Phi}\bigg(h(t) \cdot \widetilde{Sep}\big(T_{HC}(\widetilde{F})\big)\bigg),$$

where $h(t) = h(t; \beta, r, \theta, a, \Omega_p, p)$ satisfies $h(t) = 1/2 + o(1)$ when $r < \beta$ and $h(t) = L_p$ when $r \geq \beta$. Recall that in both cases, $\widetilde{Sep}(T_{ideal}) = L_p\widetilde{Sep}(T_{HC}(\widetilde{F})) = L_p p^{(1-\theta)/2-\delta(\beta,r)}$, where the exponent $(1-\theta)/2 - \delta(\beta, r)$ is strictly positive by the assumption of $r > \rho^*_\theta(\beta)$. Therefore, if (β, r) fall in Region of Possibility and if we set t as either of the two ideal thresholds, then $L_t(X, \Omega)$ not only gives successful classification, but the classification error converges to 0 very fast.

3. Classification by HCT. In the preceding section, we have been focused on two ideal thresholds. In this section, we study the empirical quantities, and characterize the stochastic fluctuation of HCT and Sep defined in (2.2). We conclude the section by proving Theorems 1.2-1.3. The main results in this section are new, even in the idealized case where $\Omega = I_p$.

3.1. *Stochastic control on the HC functional.* Recall that

$$HC(t, \bar{F}_p) = \sqrt{p}[\bar{F}_p(t) - \bar{\Psi}(t)]/\sqrt{\bar{F}_p(t)(1-\bar{F}_p(t))}.$$

When $\bar{F}_p(t) = 0$, the above is not well-defined, and we modify the definition slightly by replacing $\bar{F}_p(t)$ with $1/p$. The change does not affect the proof of the results. The stochastic fluctuation of HCT comes from that of $\bar{F}_p(t)$, which consists of two components: that of estimating Ω and that of the data. This is captured in the following triangle inequality (see (2.15)-(2.16)):

$$|\bar{F}_p(t) - \widetilde{F}(t)| \leq |\widetilde{F}_p(t) - \widetilde{F}(t)| + |\bar{F}_p(t) - \widetilde{F}_p(t)|.$$

Consider $|\widetilde{F}_p(t) - \widetilde{F}(t)|$ first. The key is to study

$$\sqrt{p}\big(\widetilde{F}_p(t) - \widetilde{F}(t)\big)/\sqrt{\widetilde{F}(t)(1-\widetilde{F}(t))}.$$

When $\Omega = I_p$, this is the *standard uniform stochastic processes* [37] and much is known about its stochastic fluctuation. In the more general case where $\Omega \neq I_p$, it is usually hard to derive a tight bound on the tail probability of this processes. Fortunately, when Ω is K_p-sparse, tight bounds are possible, and the key the *separability of sparse graphs* introduced in Lemma 1.1.

Recall that $s_p^* = \sqrt{2\log(p)}$ (e.g., (1.10)). The following lemma is the direct result of Lemma 1.1 and the well-known Bennet's inequality [37], and is proved in Section 5.

LEMMA 3.1. *Fix $(\beta, r, \theta, a) \in (0,1)^4$ and consider an ARW$(\beta, r, \theta, \Omega)$ model with $\Omega \in \mathcal{M}_p^*(a, K_p)$. As $p \to \infty$, there is a constant $C > 0$ such that with probability at least $1 - o(p^{-1})$, for all t satisfying $\bar{\Psi}^{-1}(1/2) < t < s_p^*$,*

$$\sqrt{p}|\widetilde{F}_p(t) - \widetilde{F}(t)|/\sqrt{\widetilde{F}(t)(1-\widetilde{F}(t))} \leq CK_p^3\big(\log(p)\big)^{15/4}.$$

Next, consider $|\widetilde{F}_p(t) - \bar{F}_p(t)|$. Recall that $n_p = p^\theta$. By definition, if $\hat{\Omega}$ is an acceptable estimator of Ω, then there is a constant $C > 0$ such that with probability at least $1 - o(p^{-1})$,

$$\max_{\{1\leq i,j\leq p\}}\big\{\big|\hat{\Omega}(i,j) - \Omega(i,j)\big|\big\} \leq CK_p^2\sqrt{2\log(p)}\cdot p^{-\theta/2}. \tag{3.1}$$

As a result, we have the following lemma, whose proof is straightforward and thus omitted. Recall that $\hat{Z} = \hat{\Omega}Z$ and $\tilde{Z} = \Omega Z$ (e.g., (1.9) and (2.4)).

LEMMA 3.2. *For any acceptable estimator $\hat{\Omega}$, $\max_{\{1\leq j\leq p\}}\big\{|\hat{Z}(j) - \tilde{Z}(j)|\big\} \leq CK_p^3\log(p)p^{-\theta/2}$ with probability at least $1 - o(1/p)$.*

Write for short $\eta_p = CK_p^3\log(p)p^{-\theta/2}$. By Lemma 3.2, with probability at least $1 - o(1/p)$, for all $1 \leq j \leq p$, $\big|1\{|\hat{Z}(j)| \geq t\} - 1\{|\tilde{Z}(j)| \geq t\}\big| \leq 1\{t - \eta_p \leq |\tilde{Z}(j)| \leq t + \eta_p\}$. As a result,

$$|\widetilde{F}_p(t) - \bar{F}_p(t)| \leq \widetilde{F}_p(t - \eta_p) - \widetilde{F}_p(t + \eta_p),$$

where we note that heuristically,

$$\widetilde{F}_p(t - \eta_p) - \widetilde{F}_p(t + \eta_p) \approx \widetilde{F}(t - \eta_p) - \widetilde{F}(t + \eta_p) \approx 2\eta_p|\widetilde{F}'(t)|.$$

Combining these, with probability at least $1 - o(1/p)$, for any $t > \bar{\Psi}^{-1}(\frac{1}{2})$,

$$\frac{\sqrt{p}|\widetilde{F}_p(t) - \bar{F}_p(t)|}{\sqrt{\widetilde{F}(t)(1-\widetilde{F}(t))}} \leq 2\sqrt{2p}\eta_p|\widetilde{F}'(t)|/\sqrt{\widetilde{F}(t)} = 2\sqrt{2}p^{(1-\theta)/2}|\widetilde{F}'(t)|/\sqrt{\widetilde{F}(t)}.$$

Recall $s_p^* = \sqrt{2\log(p)}$. The above heuristic is captured in the following lemma, which is proved in Section 5.

LEMMA 3.3. *Fix $(\beta, r, \theta, a) \in (0,1)^4$. In the $ARW(\beta, r, \theta, \Omega)$ model with $\Omega \in \mathcal{M}_p^*(a, K_p)$, there exists a constant $C > 0$ such that with probability at least $1 - o(1/p)$, for all t such that $\bar{\Psi}^{-1}(\frac{1}{2}) < t < s_p^*$,*

$$\sqrt{p}|\bar{F}_p(t) - \widetilde{F}_p(t)| \cdot [\widetilde{F}(t)(1 - \widetilde{F}(t))]^{-1/2} \le L_p \max\{(p^{(1-\theta)}\widetilde{F}(t))^{1/2}, 1\}.$$

Combining Lemmas 3.1 and 3.3, the following theorem follows directly.

THEOREM 3.1. *Fix $(\beta, r, \theta, a) \in (0,1)^4$. In the $ARW(\beta, r, \theta, \Omega)$ model with $\Omega \in \mathcal{M}_p^*(a, K_p)$, as $p \to \infty$, with probability at least $1 - o(p^{-1})$,*

$$\left|HC(t, \bar{F}_p) - HC(t, \widetilde{F})\right| \le L_p[(p^{1-\theta}\widetilde{F}(t))^{1/2} + 1], \qquad \forall\, \bar{\Psi}^{-1}(\frac{1}{2}) < t < s_p^*.$$

By Theorem 3.1, in order for $|T_{HC}(\bar{F}_p) - T_{HC}(\widetilde{F})|$ to be small, we must have that for all t in the vicinity of $T_{HC}(\widetilde{F})$,

$$L_p[(p^{1-\theta}\widetilde{F}(t))^{1/2} + 1] \ll HC(t, \widetilde{F}).$$

When $\theta > 1/2$, this holds for all (β, r) in Region of Possibility. When $\theta \le 1/2$, this might not hold for all (β, r) in this region, as the estimation error of $\hat{\Omega}$ is simply too large. This explains why we need to restrict HCT to be no less than $\tilde{s}_{p,n}^*$ as in (1.10). This also explains that why we need Conditions (a)-(b) in Theorem 1.3, but we don't need such conditions in Theorem 1.2 and Corollary 1.1.

In the $ARW(\beta, r, \theta, \Omega)$ model, $n_p = p^\theta$. Therefore,

$$\tilde{s}_{p,n}^* = s_p(\theta), \qquad \text{if we let } s_p(\theta) = \sqrt{2\max\{(1-2\theta), 0\}\log(p)};$$

see (1.10). Accordingly, the HCT defined in (1.11) can be rewritten as

$$t_p^{HC} = \begin{cases} T_{HC}(\bar{F}_p), & \text{if } s_p(\theta) \le T_{HC}(\bar{F}_p) \le s_p^*, \\ s_p(\theta), & \text{if } T_{HC}(\bar{F}_p) < s_p(\theta), \\ s_p^*, & \text{if } T_{HC}(\bar{F}_p) > s_p^*. \end{cases}$$

The main result in this section is as follows.

THEOREM 3.2. *Fix $(\beta, r, \theta, a) \in (0,1)^4$ such that $(1-\theta)/2 < \beta < 1 - \theta$ and $r > \rho_\theta^*(\beta)$. In the $ARW(\beta, r, \theta, \Omega)$ model with $\Omega \in \mathcal{M}_p^*(a, K_p)$,*

1) *If $\theta > \frac{1}{2}$, then as $p \to \infty$, there are positive constants $c_2 = c_2(\beta, r, a, \theta)$ and $d_0 = d_0(\beta, r, a, \theta)$ such that with probability at least $1 - o(1/p)$, $|t_p^{HC} - T_{ideal}(\epsilon_p, \tau_p, \Omega)| \le L_p p^{-c_2}$ when $r < \beta$, and $t_p^{HC} \in [\sqrt{2\beta\log p} - \Delta_1, \tau_p)$ when $r \ge \beta$, where $\Delta_1 = d_0\log(\log(p))/\sqrt{\log(p)}$.*

2) If $0 < \theta \le \frac{1}{2}$ and (β, r, θ) satisfy the conditions in Theorem 1.3, then with probability at least $1 - o(1/p)$, $|t_p^{HC} - T_{ideal}(\epsilon_p, \tau_p, \Omega)| \le L_p p^{-c_3}$ for some constant $c_3 = c_3(\beta, r, a) > 0$ when $r < \beta$, and $t_p^{HC} \in [\sqrt{2\beta \log p} - \Delta_1, \tau_p)$ for $\Delta_1 = d_1 \log(\log(p))/\sqrt{\log p}$ when $r \ge \beta$, where $d_1 = d_1(\beta, r, a) > 0$ is a constant.

3.2. *Stochastic fluctuation of Fisher's separation.* Similarly, the stochastic fluctuation of $Sep(t, \hat{Z}, \mu, \hat{\Omega})$ contains two parts: that from $\tilde{Z} = \Omega Z$, and that from the estimation $\hat{\Omega}$. In detail,

$$|Sep(t, \hat{Z}, \mu, \hat{\Omega}) - \widetilde{Sep}(t, \epsilon_p, \tau_p, \Omega)| \le 2 \cdot (I + II),$$

where $I = \frac{1}{2}|Sep(t, \tilde{Z}, \mu, \Omega) - \widetilde{Sep}(t, \epsilon_p, \tau_p, \Omega)|$ and $II = \frac{1}{2}|Sep(t, \hat{Z}, \mu, \hat{\Omega}) - \widetilde{Sep}(t, \tilde{Z}, \mu, \Omega)|$.

Consider I first. Recall that

$$Sep(t, \tilde{Z}, \mu, \Omega) = 2M_p(t, \tilde{Z}, \mu, \Omega))/\sqrt{V_p(t, \tilde{Z}, \Omega)}.$$

Heuristically, $M_p(t, \tilde{Z}, \mu, \Omega) = m_p(t, \epsilon_p, \tau_p, \Omega) + O_p(\sqrt{m_p(t, \epsilon_p, \tau_p, \Omega)})$ and $V_p(t, \tilde{Z}, \mu, \Omega) = v_p(t, \epsilon_p, \tau_p, \Omega) + O_p(\sqrt{v_p(t, \epsilon_p, \tau_p, \Omega)})$; see (2.6). Combining these with the definitions, we expect that

(3.2)
$$Sep(t, \tilde{Z}, \mu, \Omega) = \widetilde{Sep}(t, \epsilon_p, \tau_p, \Omega)\big[1 + O_p\big(\frac{1}{\sqrt{m_p(t, \epsilon_p, \tau_p, \Omega)}} + \frac{1}{\sqrt{v_p(t, \epsilon_p, \tau_p, \Omega)}}\big)\big],$$

where in the bracket, the second term is much smaller than 1. This is elaborated in the following lemma which is proved in Section 5. In detail, let $q(t) = q(t; \beta, r, \theta, \Omega_p, p)$ satisfy that $q(t) = p^{(1-\theta)/2 - \max\{4\beta - 2r, 3\beta + r\}/4}$ if $r < \beta$ and $q(t) = 0$ if $r \ge \beta$.

LEMMA 3.4. *Fix $(\beta, r, \theta, a) \in (0, 1)^4$ such that $r > \rho_\theta^*(\beta)$ and $(1-\theta)/2 < \beta < (1-\theta)$. In the ARW$(\beta, r, \theta, \Omega)$ model with $\Omega \in \widetilde{\mathcal{M}}_p^*(a, b, K_p)$, as $p \to \infty$, with probability at least $1 - o(1/p)$,*

$$\sup_{\{t>0\}} |Sep(t, \tilde{Z}, \mu, \Omega) - \widetilde{Sep}(t, \epsilon_p, \tau_p, \Omega)| \le L_p[q(t) + p^{-\theta/2}].$$

When $r < \beta$, the condition on Ω can be relaxed to that of $\Omega \in \mathcal{M}_p^*(a, K_p)$.

Next, we consider II. The following lemma, which is proved in Section 5, characterizes the order of II.

LEMMA 3.5. *Under the same conditions as in Lemma 3.4, as $p \to \infty$, with probability at least $1 - o(1/p)$, for all t such that $s_p(\theta) < t < s_p^*$, $|Sep(t, \hat{Z}, \mu, \hat{\Omega}) - Sep(t, \tilde{Z}, \mu, \Omega)| \le L_p[p^{-\theta}(p\widetilde{F}(t))^{1/2} + q(t) + p^{-\theta/2}]$. When $r < \beta$, the condition on Ω can be relaxed to that of $\Omega \in \mathcal{M}_p^*(a, K_p)$.*

Combining Lemmas 3.4–3.5, we have the following theorem, which is parallel to Theorem 3.1 and is proved in Section 5.

THEOREM 3.3. *Under the same conditions as in Lemma 3.4, as $p \to \infty$, with probability at least $1 - o(p^{-1})$, for all t such that $s_p(\theta) < t < s_p^*$,*

$$\Big|Sep(t, \hat{Z}, \mu, \hat{\Omega}) - \widetilde{Sep}(t, \epsilon_p, \tau_p, \Omega)\Big| \le L_p[p^{-\theta}(p\widetilde{F}(t))^{1/2} + p^{-\theta/2} + q(t)].$$

When $r < \beta$, the condition on Ω can be relaxed to that of $\Omega \in \mathcal{M}_p^*(a, K_p)$.

3.3. *Proof of Theorems 1.2–1.3.* We are now ready to prove Theorems 1.2–1.3, where Ω is assumed as known and unknown, respectively. The proofs are similar, so we only show Theorem 1.3. Consider $L_{HC}(X, \hat{\Omega})$, where $\hat{\Omega}$ is an acceptable estimator. The misclassification error is

$$P\big(Y \cdot L_{HC}(X, \hat{\Omega}) < 0\big) = E_{\epsilon_p, \tau_p} E\bigg[\bar{\Phi}\big(\frac{1}{2} Sep(t_p^{HC}, \hat{Z}, \mu, \hat{\Omega})\big)\bigg]. \tag{3.3}$$

We now prove for the case of $r < \beta$ and $r \ge \beta$ separately.

In the first case, we note that $L_p[p^{-\theta}(p\widetilde{F}(t))^{1/2} + p^{-\theta/2}] \le L_p p^{\min\{0, \frac{1}{2}-\theta\}}$ for $s_p(\theta) < t < s_p^*$. Write $T_{ideal} = T_{ideal}(\epsilon_p, \tau_p, \Omega)$ and $\widetilde{Sep}(t) = \widetilde{Sep}(t, \epsilon_p, \tau_p, \Omega)$ for short as before. By Theorem 3.3, with probability $1 - o(1/p)$,

$$|Sep(t_p^{HC}, \hat{Z}, \mu, \hat{\Omega}) - \widetilde{Sep}(t_p^{HC})| \le L_p[p^{\min\{0, \frac{1}{2}-\theta\}} + p^{\frac{1-\theta}{2} - \max\{\beta - \frac{r}{2}, \frac{3\beta+r}{4}\}}]. \tag{3.4}$$

At the same time, by Theorem 3.2, with probability $1 - o(1/p)$, $|t_p^{HC} - T_{ideal}|$ is algebraically small. Note that $\widetilde{Sep}(t)$ is a non-stochastic function. By Taylor expansion and Lemma 2.1,

$$\widetilde{Sep}(t_p^{HC}) = (1 + o(1))\widetilde{Sep}(T_{ideal}) = L_p p^{\frac{1-\theta}{2} - \delta(\beta, r)}, \tag{3.5}$$

where $\delta(\beta, r)$ is as in (2.10). By definitions, $\max\{4\beta - 2r, 3\beta + r\}/4 > \delta(\beta, r)$. Inserting (3.3)-(3.5) into (3.3) gives

$$P\big(Y \cdot L_{HC}(X, \hat{\Omega}) < 0\big) = \big(1 + o(1/p)\big)\bar{\Phi}\Big(L_p p^{\frac{1-\theta}{2} - \delta(\beta, r)}\Big) + o(1/p), \tag{3.6}$$

and the claim follows since $(1 - \theta)/2 - \delta(\beta, r) > 0$.

In the second case, $\sqrt{2\beta\log p} \lesssim t_p^{HC} \lesssim \sqrt{2r\log p}$ with probability at least $1-o(1/p)$. Combining this with Theorem 3.3, with probability at least $1-o(1/p)$,

$$|Sep(t_p^{HC}, \hat{Z}, \mu, \hat{\Omega}) - \widetilde{Sep}(t_p^{HC})| \le L_p p^{\min\{0,\frac{1}{2}-\theta\}}. \tag{3.7}$$

At the same time, by similar argument as that of the proof of Theorem 2.2,

$$2\tau_p K_p^{-1} p^{(1-\theta-\beta)/2} \lesssim \widetilde{Sep}(t_p^{HC}) \le \widetilde{Sep}(T_{ideal}) = L_p p^{(1-\theta-\beta)/2}.$$

Combining this with (3.3) and (3.7) gives

$$P\big(Y\cdot L_{HC}(X,\hat{\Omega}) < 0\big) = (1+o(1/p))\bar{\Phi}\Big(\frac{1}{2}L_p p^{(1-\theta)/2-\delta(\beta,r)}\Big) + o(1/p), \tag{3.8}$$

and the claim follows since $\frac{1-\theta}{2} - \delta(\beta,r) > 0$. This proves Theorem 1.3.

We conclude this section by a remark on the convergence rate. At the end of Section 2, we show that the 'ideal' classifier $L_t(X,\Omega)$ have very fast convergence rate with t being either the ideal threshold or the ideal HCT. In comparison, the convergence rate of $L_{HC}(X,\hat{\Omega})$ is unfortunately much slower (but is still algebraically fast). To explain this, we note that the rate of convergence of t_p^{HC} to $T_{HC}(\widetilde{F})$ and the rate of convergence of $\hat{\Omega}$ to Ω are both algebraically fast; if these convergence rates can be improved, then the misclassification error rate of $L_{HC}(X,\hat{\Omega})$ can be improved as well.

4. Simulations. We have conducted a small-scale numerical study. The idea is to select a few sets of representative parameters for experiments, and compare the performance of HCT classifier (HCT) with three other methods: ordinary HCT (oHCT), pseudo HCT (pHCT), and CVT. All these methods are very similar to HCT, except for that (a) in pHCT, we assume Ω is known to us, (b) in CVT, we set the threshold of IT by a 5-fold cross validation, and (c) in oHCT, we pretend Σ is diagonal, and estimate Ω accordingly. Note that CVT reduces to PAM [39] if we do not utilize the correlation structure; see more discussion in [15].

4.1. *Estimating* Ω. For some of the procedures, we need to estimate Ω. We use Bickel and Levina's Thresholding (BLT) procedure [4]. Alternatively, one could use the glasso [21] or the CLIME [9]. But since the main goal is to investigate the performance of HCT, we do not include glasso and CLIME in the study: if HCT performs well with Ω estimated by BLT, we expect it to perform even better if Ω is estimated more accurately.

At the same time, each of these methods can be improved numerically with an additional *re-fitting* stage. Take the BLT for example. For the training data $\{(X_i,Y_i)\}_{i=1}^n$, let $\bar{X} = \frac{1}{n}\sum_{i=1}^n Y_i X_i$, and let $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (Y_i X_i -$

$\bar{X})'(Y_iX_i - \bar{X})$ be the empirical covariance matrix. BLT starts by obtaining an estimate of Σ using thresholding:

$$\Sigma^*(i,j) = \hat{\Sigma}(i,j)1\{|\hat{\Sigma}(i,j)| \geq \eta\}, \qquad 1 \leq i,j \leq p, \tag{4.1}$$

and then estimate Ω by $\hat{\Omega}^{**} = (\Sigma^*)^{-1}$. Here, $\eta > 0$ is a tuning parameter.

We propose the following refitting stage to improve the estimator. Fixing a tuning parameter $\zeta > 0$, we further improve $\hat{\Omega}^{**}$ via coordinate-wise thresholding and call the resultant estimator $\hat{\Omega}^*$:

$$\hat{\Omega}^*(i,j) = \hat{\Omega}^{**}(i,j)1\{|\hat{\Omega}^{**}(i,j)| \geq \zeta\}. \tag{4.2}$$

For each $1 \leq i \leq p$, let $S_i = \{1 \leq j \leq p : \hat{\Omega}^*(i,j) \neq 0\}$, and let A_i be the sub-matrix of $\hat{\Sigma}$ formed by restricting the rows/columns of $\hat{\Sigma}$ to S_i. Denote the final estimate of Ω by $\hat{\Omega} = [\omega_1, \omega_2, \ldots, \omega_p]$. We define ω_i as follows. Write $S_i = \{j_1, j_2, \ldots, j_k\}$, where $k = |S_i|$. Let e_i be the $p \times 1$ vector such that $e_i(j) = 1\{i = j\}$, $1 \leq j \leq p$, and let ξ_i be the $k \times 1$ vector formed by restricting the rows of e_i to S_i. Define $\eta_i = A_i^{-1}\xi_i$. We let $\omega_i(j_\ell) = \eta_i(\ell)$, $1 \leq \ell \leq k$, and let $\omega_i(j) = 0$ if $j \notin S_i$.

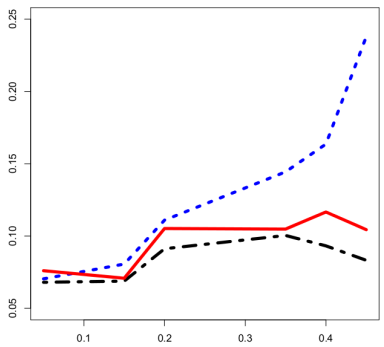

FIG 1. *Comparison of classification errors by HCT (solid), oHCT (dashed) and pHCT (dash-dotted). The x-axis is a, and the y-axis is the classification error (Experiment 1a).*

4.2. *Numerical experiments.* Fix $(p, n, \epsilon_p, H_p, \Omega)$ and an integer m, each simulation experiment contains the following main steps.

1. Generate a $p \times 1$ vector μ according to $(\sqrt{n}\mu(j)) \overset{iid}{\sim} (1-\epsilon_p)\nu_0 + \epsilon_p H_p$.
2. Generate training data (X_i, Y_i), $1 \leq i \leq n$, by letting $Y_i = 1$ for $i \leq n/2$ and $Y_i = -1$ for $i > n/2$, and $X_i \sim N(Y_i \cdot \mu, \Omega^{-1})$.

3. Generate m test vectors, each of which has the form of $X \sim N(Y \cdot \mu, \Omega^{-1})$, where $Y = \pm 1$ with equal probabilities.
4. Use the training data to build all four classifiers, apply them to the test set, and then record the test errors.

When we need to estimate Ω, we use BLT with the aforementioned refitting stage. The study contains three different experiments, which we now discuss separately.

Experiment 1. In this experiment, we compare HCT with oHCT and pHCT. The experiment contains three sub-experiments 1a, 1b and 1c.

In Experiment 1a, we fix $(p, n, \epsilon_p, \tau_p, m) = (3000, 2000, 0.1, 4, 500)$, and let H_p be the point mass at τ_p. Also, we choose Ω to be the tridiagonal matrix

$$\Omega(i,j) = 1\{i = j\} + a \cdot 1\{|i - j| = 1\}, \qquad 1 \le i, j \le p, \tag{4.3}$$

where a takes values from $\{.05, .15, .2, .35, .4, .45\}$. The results are reported in Figure 1. The tuning parameter η in (4.1), which varies with the values of a, n and p, is calculated from trials of comparing $(\Sigma^*)^{-1}$ with the true Ω. The tuning parameter ζ in (4.2), which also varies with the values of a, n and p, is chosen so that there are only k nonzero coordinates in each row of $\hat{\Omega}^*$ after thresholding of $\hat{\Omega}^{**}$. We let $k = 2, 3$ if Ω is tridiagonal and $k = 4, 5$ if Ω is five-diagonal (see experiments below). In this experiment, η is set accordingly from $\{.1, .1, .15, .15, .2, .25\}$ and ζ is from $\{.05, .1, .1, .2, .25, .3\}$. The results suggest that HCT outperforms oHCT, but is slightly inferior to pHCT since we have to pay a price for estimating Ω. As a increases, the correlation structure becomes increasingly influential, so the advantage of HCT over oHCT becomes increasingly prominent (but differences between HCT and pHCT remain almost the same).

In Experiment 1b, for various $(p, n, \epsilon_p, \tau_p)$, we choose $m = 500$ and let Ω be either of the following tridiagonal matrix or five-diagonal matrix. In the first case, Ω is a $p \times p$ tridiagonal matrix with 1 on the diagonal and a on the off-diagonal. In the second case, Ω is a $p \times p$ five-diagonal matrix with 1 on the diagonal, a_1 on the first off-diagonal, and a_2 on the second off-diagonal. Experiment 1c uses a very similar setting, except that we take H_p as the uniform distribution over $[\tau_p - 0.5, \tau_p + 0.5]$. We select ζ and η similarly as in experiment 1a. The results based on 5 repetitions for Experiment 1b-1c are reported in Table 1, which suggest that HCT outperforms oHCT and that pHCT slightly outperforms HCT.

Experiment 2. In this experiment, we compare the pHCT with the CVT assuming Ω is known (the case Ω is unknown is discussed in Experiment 3). Experiment 2 contains two sub-experiments, 2a and 2b.

	$n=1000, p=2000$	$n=2000, p=3000$	$n=2000,\ p=3000$
	$a=.05, \epsilon_p=.1, \tau_p=4$	$a=.45,\ \epsilon_p=.2, \tau_p=3$	$a_1=.45,\ a_2=.2\ ,\ \epsilon_p=.1, \tau_p=4$
oHCT	0.054	0.2616	0.17
pHCT	0.0448	0.058	0.098
HCT	0.052	0.061	0.0992
	$n=500, p=1000$	$n=2000,\ p=3000$	$n=2000,\ p=3000$
	$a=.05, \epsilon_p=0.1, \tau_p=4$	$a=.45,\ \epsilon_p=.05, \tau_p=5$	$a_1=.35,\ a_2=.2,\ \epsilon_p=.1,\ \tau_p=4$
oHCT	0.0536	0.2268	0.1332
pHCT	0.046	0.1284	0.0912
HCT	0.0524	0.1344	0.1252
	$n=1000, p=2000$	$n=2000, p=3000$	$n=2000,\ p=3000$
	$H_p=U(3.5, 4.5)$	$H_p=U(2.5, 3.5)$	$H_p=U(3.5, 4.5)$
	$a=.05, \epsilon_p=.1$	$a=.45, \epsilon_p=.2, \tau_p=3$	$a_1=.45, a_2=.2, \epsilon_p=.1, \tau_p=4$
oHCT	0.052	0.2816	0.1472
pHCT	0.046	0.0704	0.0840
HCT	0.044	0.0716	0.0891

TABLE 1

Classification errors by HCT, oHCT and pHCT. Ω is tridiagonal (left two columns) or five-diagonal matrix (right column). Rows 1-2: Experiment 1b. Row 3: Experiment 1c.

In Experiment 2a, we consider 6 different combinations of $(p, n, \epsilon_p, \tau_p)$ with $m = 500$, and let Ω be the tridiagonal matrix as in (4.3) with $a = 0.2$. Averages of the selected thresholds and classification errors across different replications are reported in Table 2. The results suggest that the threshold choices by HC and cross validations are considerably different, with the former being more accurate and more stable. Note that HCT is also computationally much more efficient than the CVT.

	Threshold	Error	Threshold	Error	Threshold	Error
pHCT	1.9	0.05	2.16	0.002	1.99	0
CVT	2.5	0.08	1	0.018	1	0
pHCT	2.39	0.18	2.06	0.10	2.13	0.02
CVT	1.9	0.224	2.00	0.14	1.1	0.09

TABLE 2

Comparison of thresholds (Column 2, 4, 6) and classification errors (Column 3,5, 7) by pHCT and CVT. $(p, \tau_p) = (3000, 1.8)$, and $\epsilon_p = 0.1$ (top) and 0.05 (bottom). Left to right: $n = 100, 50, 20$ (Experiment 2a).

In Experiment 2b, we set $(p, \epsilon_p, m) = (3000, 0.05, 500)$, $n \in \{20, 40\}$, and let Ω be the same as in Experiment 2a. We let τ_p range from 1 to 2.5 with an increment of 0.1. The classification errors by pHCT and CVT are in Figure 2, where a similar conclusion can be drawn as that in Experiment 2a.

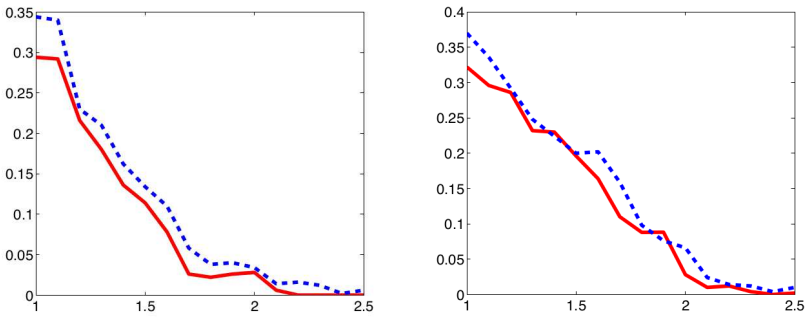

FIG 2. *Classification errors of pHCT (solid) and CVT (dashed) for* $n = 20$ *(left) and* 40 *(right) and various* τ_p *(x-axis) (Experiment 2b).*

Experiment 3. We compare the performance of HCT with CVT for the case where Ω is unknown and needs to be estimated. Note that for small n (say, less than 500) we might not have reasonable accuracy on estimating Ω using BLT. For small p, say 100-300, the CVT is computationally very slow and it is very likely that the refitting procedure for BLT would not have decent performance. We take $(p, n, \epsilon_p) = (5,000, 500, .1)$ and let Ω be the block diagonal matrix consisting 10 diagonal blocks, each is a big five-diagonal matrix $C = C_{500,500}(a_1, a_2)$, where $C(i, j) = 1\{i = j\} + a_1 \cdot 1\{|i - j| = 1\} + a_2 \cdot 1\{|i - j| = 2\}$, $1 \le i, j \le 500$, and $a_1 = .45$, $a_2 = .1$. We let τ_p range from 1 to 3 with an increment of 0.2. The tuning parameter ζ and η are set in the similar way as in Experiment 1. The results are reported in Figure 3. Due to high computational cost, we only conduct $m = 6$ repetitions, so the results are a bit noisy. Still, it is seen that HCT outperforms CVT.

In summary, for a reasonably large sample size n, HCT outperforms oHCT and is only slightly inferior to pHCT. The reason we need a relatively large n is mainly due to that we need to estimate Ω. The relative performance of pHCT, HCT, and oHCT is intuitive, since pHCT utilizes the true correlation structure among the features, HCT estimates the correlation structure, while oHCT ignores it. The comparisons of pHCT with CVT in Experiments 2a-2b suggest that if Ω is known, then HCT dominates CVT. Experiment 3 shows that when p is several times larger than n (e.g., 10 times larger), HCT has smaller classification errors than CVT does, and the precision matrix Ω can be estimated reasonably well.

For larger p, the advantages of the HCT are even more prominent than those considered here. We skip the comparisons for larger p due to high computational cost, which mainly comes from the BLT procedure (we must

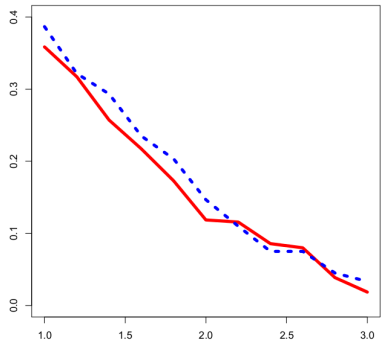

FIG 3. *Classification errors by HCT (solid) and CVT (dashed) for various* τ_p *(x-axis)(Experiment 3).*

run the algorithm many times to select a good tuning parameter η). In the future, if we could find a more efficient method for estimating Ω, then HCT will be both more effective and more convenient to use for large p.

5. Proofs. In this section, we prove all key theorems and lemmas in the order they appear (except for Theorem 1.2-1.3 which are proved in Section 3.3). Secondary lemmas are proved in Section 6.

5.1. *Proof of Theorem 1.1.* For short, write $n = n_p$. Recall that the training samples are $X_i \sim N(Y_i\mu, \Omega^{-1})$, $1 \le i \le n$, where $Y_i \in \{-1, 1\}$ are given. Consider an (independent) test sample $X \sim N(Y \cdot \mu, \Omega^{-1})$, where $Y = \pm 1$ with equal probabilities. Let $f_{\pm 1}$ be the joint of density of $(X_1, \ldots, X_n, X)$ in the case where $Y = 1$ and $Y = -1$, respectively, and let $H(f, g)$ be the Hellinger distance between two density functions f and g. To show the claim, it is sufficient to show $H(f_1, f_{-1}) \to 0$ as $p \to \infty$, uniformly for all $\Omega \in \mathcal{M}_p^*(a, K_p)$. Let f_0 be the joint density of $(X_1, \ldots, X_n, X)$ in the case where $X \sim N(0, \Omega^{-1})$ (but the distributions of X_i remain the same). By triangle inequality and symmetry, $H(f_1, f_{-1}) \le H(f_1, f_0) + H(f_{-1}, f_0) = 2H(f_1, f_0)$. Therefore, it is sufficient to show

$$H(f_1, f_0) \to 0. \tag{5.1}$$

Since Ω is a K_p-sparse correlation matrix, by Lemma 1.1, there is a permutation matrix P and an integer $M_p = M_p(\Omega, K_p)$ such that $M_p \le CK_p \log(p)$

and

$$P\Omega P' = \begin{pmatrix} \tilde{\Omega}_{11} & \dots & \tilde{\Omega}_{1M_p} \\ \dots & \dots & \dots \\ \tilde{\Omega}_{M_p1} & \dots & \tilde{\Omega}_{M_pM_p} \end{pmatrix}, \tag{5.2}$$

where on the diagonal, $\tilde{\Omega}_{11}, \dots, \tilde{\Omega}_{M_pM_p}$ are identity matrices. Since permuting the coordinates of $X_1, X_2, \dots, X$ simultaneously does not change the Hellinger distance $H(f_1, f_0)$, we assume $P = I_p$ for simplicity.

Now, corresponding to the partition of Ω in (5.2), we partition the mean-vector μ as $\mu = ((\mu^{(1)})', \dots, (\mu^{(M_p)})')'$. For $0 \le m \le M_p$, let P_m be the projection matrix such that $P_m\mu = ((\mu^{(1)})', \dots, (\mu^{(m)})', 0, \dots, 0)'$, where generically, 0 denotes a row vector of zeros, and let $f^{(m)}$ be the joint density of $(X_1, \dots, X_n, X)$ under the law that $X_i \sim N(Y_i\mu, \Omega^{-1})$ for all $1 \le i \le n$ and $X \sim N(P_m\mu, \Omega^{-1})$. Note that $f_0 = f^{(0)}$ and $f_1 = f^{(M_p)}$, and that by triangle inequality,

$$H(f^{(0)}, f^{(M_p)}) \le \sum_{m=1}^{M_p} H(f^{(m-1)}, f^{(m)}). \tag{5.3}$$

Recalling $M_p \le CK_p \log(p)$ and $K_p \le L_p$, (5.1) follows by Lemma 5.1 below. □

LEMMA 5.1. *There is a constant $c_0 = c_0(\beta, r, \theta) > 0$ such that for any $1 \le m \le M_p - 1$,*

$$H(f^{(m-1)}, f^{(m)}) \le L_p p^{-c_0}. \tag{5.4}$$

5.2. *Proof of Lemma 5.1.* Denote $K = K_p$, $M = M_p$, and $n = n_p$ for short. Recall that each of $X, X_1, \dots, X_n$ can be partitioned into M blocks. We simultaneously swap the first block and the m-th block of X and of each X_i, but still denote the resultant vectors by X and X_i for notational simplicity. Denote $\tilde{\nu} = \mu^{(m)}$, $\tilde{\tilde{\nu}} = ((\mu^{(1)})', \dots, (\mu^{(m-1)})', 0, \dots, 0)'$, and $\tilde{\tilde{\mu}} = ((\mu^{(1)})', (\mu^{(2)})', \dots, (\mu^{(m-1)})', (\mu^{(m+1)})', \dots, (\mu^{(M)})')'$. After the swaps, $f^{(m)}$ is the joint density of $(X_1, \dots, X_n, X)$, where the common mean vector of $X_1, \dots, X_n$ (which we still denote by μ for simplicity) is $\mu = (\tilde{\nu}', \tilde{\tilde{\mu}}')'$, the mean vector of X is $(\tilde{\nu}', \tilde{\tilde{\nu}}')'$, and the common precision matrix (still denote by Ω for simplicity) of $X_1, \dots, X_n, X$ is

$$\Omega = \begin{pmatrix} I_k & B \\ B' & D \end{pmatrix}, \tag{5.5}$$

where I_k is a $k \times k$ identity matrix with $k = k(\Omega, m)$ equals to the size of the m-th block (before the swaps) and D is a correlation matrix. Similarly, $f^{(m-1)}$ is the joint density of $(X_1, \ldots, X_n, X)$, where the laws of $X_1, \ldots, X_n, X$ are the same as that of $f^{(m)}$ except for that the mean vector of X is $(0, \tilde{\tilde{\nu}}')'$ instead.

Denote for short $f_0 = f^{(m-1)}$, $f_1 = f^{(m)}$. Since Y_i are given, we assume $Y_i = 1$ for notational simplicity. Consequently, $Z = \frac{1}{\sqrt{n}} \sum_{i=1}^n Y_i X_i$ reduces to $Z = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$. By definitions and elementary statistics, $f_0(x_1, \ldots, x_n, x) = \phi(x, \Omega) \Pi_{i=1}^n \phi(x_i, \Omega) \cdot I$, and $f_1(x_1, \ldots, x_n, x) = \phi(x, \Omega) \Pi_{i=1}^n \phi(x_i, \Omega) \cdot II$, where

$$I = \int e^{\sqrt{n}\mu'\Omega z - \frac{n}{2}\mu'\Omega\mu + (0, \tilde{\tilde{\nu}}')\Omega x - \frac{1}{2}\tilde{\tilde{\nu}}' D \tilde{\tilde{\nu}}} dF(\mu),$$

$$II = \int e^{\sqrt{n}\mu'\Omega z - \frac{n}{2}\mu'\Omega\mu + (\tilde{\nu}', \tilde{\tilde{\nu}}')\Omega x - \frac{1}{2}[\|\tilde{\nu}\|^2 + \tilde{\nu}' B \tilde{\tilde{\nu}} + \tilde{\tilde{\nu}}' D \tilde{\tilde{\nu}}]} dF(\mu),$$

and $F(\mu)$ denotes the cdf of μ. Here, x and x_i are $p \times 1$ vectors, $z = \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i$, and $\phi(x, \Omega)$ is the joint density of $N(0, \Omega^{-1})$. For $1 \le i \le k$, denote the i-th row of B in (5.5) by ω_i'. Also, write $\Omega x = (\tilde{x}, \tilde{\tilde{x}})'$ and $\Omega z = (\tilde{z}, \tilde{\tilde{z}})'$ so that the lengths of $\tilde{x}$ and $\tilde{z}$ are k. Introduce $g = g(\tilde{z}, \tilde{\tilde{\mu}})$, $h = h(\tilde{z}, \tilde{x}, \tilde{\tilde{\mu}}, \tilde{\tilde{\nu}})$, and $w = w(\tilde{\tilde{z}}, \tilde{\tilde{\mu}}, \tilde{\tilde{\nu}})$ by

$$g = \Pi_{i=1}^k \big[(1 - \epsilon_p) + \epsilon_p e^{\tau_p \tilde{z}_i - \frac{1}{2}\tau_p^2 - \sqrt{n}\tau_p(\omega_i, \tilde{\tilde{\mu}})}\big],$$

$$hg = \Pi_{i=1}^k \big[(1 - \epsilon_p) + \epsilon_p e^{\tau_p \tilde{z}_i + (\tau_p/\sqrt{n})\tilde{x}_i - \frac{1}{2}\tau_p^2 - \frac{1}{2n}\tau_p^2 - \sqrt{n}\tau_p(\omega_i, \tilde{\tilde{\mu}}) - (\tau_p/\sqrt{n})(\omega_i, \tilde{\nu})}\big],$$

and

$$w = e^{\sqrt{n}\tilde{\tilde{\mu}}'\tilde{\tilde{z}} + \tilde{\tilde{\mu}}'\tilde{\tilde{z}} - \frac{n}{2}\tilde{\tilde{\mu}}' D \tilde{\tilde{\mu}} - \frac{1}{2}\tilde{\tilde{\nu}}' D \tilde{\tilde{\nu}}}.$$

Here, we have suppressed the expressions of g, h, and w as long as there is no confusion. Since $\tilde{\nu}$ and $\tilde{\tilde{\mu}}$ are independent, by direct calculations,

$$\begin{aligned} I &= \int e^{\sqrt{n}\tilde{\mu}'\tilde{z} + \sqrt{n}\tilde{\tilde{\mu}}'\tilde{\tilde{z}} + \tilde{\tilde{\nu}}'\tilde{\tilde{x}} - \frac{n}{2}\|\tilde{\mu}\|^2 - n\tilde{\mu}' B \tilde{\tilde{\mu}} - \frac{n}{2}\tilde{\tilde{\mu}}' D \tilde{\tilde{\mu}} - \frac{1}{2}\tilde{\tilde{\nu}}' D \tilde{\tilde{\nu}}} dF(\tilde{\mu}) dF(\tilde{\tilde{\mu}}) \\ &= \int \left(\Pi_{i=1}^k \big[(1 - \epsilon_p) + \epsilon_p e^{\tau_p \tilde{z}_i - \frac{1}{2}\tau_p^2 - \sqrt{n}\tau_p(\omega_i, \tilde{\tilde{\mu}})}\big] \right) e^{\sqrt{n}\tilde{\tilde{\mu}}'\tilde{\tilde{z}} + \tilde{\tilde{\mu}}'\tilde{\tilde{z}} - \frac{n}{2}\tilde{\tilde{\mu}}' D \tilde{\tilde{\mu}} - \frac{1}{2}\tilde{\tilde{\nu}}' D \tilde{\tilde{\nu}}} dF(\tilde{\tilde{\mu}}), \end{aligned}$$

which, by the definitions, implies that $I = \int g w dF(\tilde{\tilde{\mu}})$. Similarly, $II = \int hgw dF(\tilde{\tilde{\mu}})$.

Let $A(f_0, f_1)$ and $H(f_0, f_1)$ be the Hellinger affinity and the Hellinger distance between f_0 and f_1, respectively. It is well-known that there is a universal constant $C > 0$ such that

$$|1 - A(f_0, f_1)| \le C \cdot H(f_0, f_1). \tag{5.6}$$

Let E_0 be the expectation under the law that $X_1, \ldots, X_n, X$ are iid from $N(0, \Omega^{-1})$. By Hölder inequality, $H(f_0, f_1) \leq E_0[(\int (h-1) g w dF(\tilde{\tilde{\mu}}))^2 / (\int g w dF(\tilde{\tilde{\mu}}))] \leq E_0[\int (h-1)^2 g w dF(\tilde{\tilde{\mu}})]$. Since $E_0[\int h g w dF(\tilde{\tilde{\mu}})] = 1$ and $E_0[\int g w dF(\tilde{\tilde{\mu}})] = 1$, it is seen

$$H(f_0, f_1) \leq E_0[\int h^2 g w dF(\tilde{\tilde{\mu}})] - 1. \tag{5.7}$$

Note that $h^2 g$ does not depend on $\tilde{\tilde{x}}$ and $\tilde{\tilde{z}}$, and that $(\tilde{\tilde{x}}|\tilde{x})$ is independent of $(\tilde{\tilde{z}}|\tilde{z})$ and $(\tilde{\tilde{x}}|\tilde{x}) \sim N(B'\tilde{x}, D - B'B)$, $(\tilde{\tilde{z}}|\tilde{z}) \sim N(B'\tilde{z}, D - B'B)$. It follows that $E\big[w\big|(\tilde{x}, \tilde{z})\big] = \exp(\sqrt{n}\tilde{\tilde{\mu}}' B'\tilde{z} - \frac{n}{2}\tilde{\tilde{\mu}}' B'B\tilde{\tilde{\mu}} + \tilde{\tilde{\nu}}' B'\tilde{x} - \frac{1}{2}\tilde{\tilde{\nu}}' B'B\tilde{\tilde{\nu}})$. Denote the right hand side by $v = v(\tilde{x}, \tilde{z}, \tilde{\tilde{\mu}}, \tilde{\tilde{\nu}})$. It follows that $E_0[\int h^2 g w dF(\tilde{\tilde{\mu}})] = E_0[\int h^2 g v dF(\tilde{\tilde{\mu}})]$. Combining this with (5.6)-(5.7) gives

$$|1 - A(f_0, f_1)| \leq C\big(E_0[\int h^2 g v dF(\tilde{\tilde{\mu}})] - 1\big) \equiv C(IV - 1). \tag{5.8}$$

We now evaluate IV. For simplicity, we assume H_p is a point mass at τ_p; the proof for general cases is similar since the support of H_p is contained in $[-\tau_p, \tau_p]$, but we need to have an extra layer of integral so the expression is much more cumbersome. Denote for short $a_i = (1 - \epsilon_p)$ and $b_i = 1 - \epsilon_p + \epsilon_p \exp(\tau_p \tilde{z}_i - \frac{\tau_p^2}{2} - \sqrt{n}\tau_p(\omega_i, \tilde{\tilde{\mu}}))$, $1 \leq i \leq k$. By direct calculations,

$$IV = E_0\bigg[\int \Pi_{i=1}^k \bigg(e^{\sqrt{n}(\omega_i, \tilde{\tilde{\mu}})\tilde{z}_i - \frac{n}{2}(\omega_i, \tilde{\tilde{\mu}})^2} \frac{\big[a_i + b_i e^{\frac{\tau_p}{\sqrt{n}}\tilde{x}_i - \frac{\tau_p^2}{2n} - \frac{\tau_p}{\sqrt{n}}(\omega_i, \tilde{\nu})}\big]^2}{a_i + b_i} e^{(\tilde{\tilde{\nu}}, \omega_i)\tilde{x}_i - \frac{1}{2}(\tilde{\tilde{\nu}}, \omega_i)^2} \bigg) dF(\tilde{\tilde{\mu}})\bigg]. \tag{5.9}$$

Recall that $\tilde{x}$ and $\tilde{z}$ are independent normal vector with I_k as the covariance matrix. It follows

$$E_0\big[(a_i + b_i e^{\frac{\tau_p}{\sqrt{n}}\tilde{x}_i - \frac{1}{2n}\tau_p^2 - \frac{\tau_p}{\sqrt{n}}(\omega_i, \tilde{\nu})})^2 e^{(\omega_i, \tilde{\tilde{\nu}})\tilde{x}_i - \frac{1}{2}(\omega_i, \tilde{\tilde{\nu}})^2}\big] = (a_i + b_i)^2 + (e^{\frac{\tau_p^2}{n}} - 1) b_i^2. \tag{5.10}$$

Denote for short $\sqrt{n}(\omega_i, \tilde{\tilde{\mu}}) = d_i \tau_p$. By definitions and direct calculations,

$$E_0\big[e^{\sqrt{n}(\omega_i, \tilde{\tilde{\mu}})\tilde{z}_i - \frac{n}{2}(\omega_i, \tilde{\tilde{\mu}})^2}(a_i + b_i)\big] = 1, \tag{5.11}$$

and

$$E_0\big[e^{\sqrt{n}(\omega_i, \tilde{\tilde{\mu}})\tilde{z}_i - \frac{n}{2}(\omega_i, \tilde{\tilde{\mu}})^2} \frac{b_i^2}{a_i + b_i}\big] = \epsilon_p^2 e^{\tau_p^2} \cdot E\bigg[\frac{e^{(2+d_i)\tau_p z_i - (2+d_i)^2 \tau_p^2/2}}{(1 - \epsilon_p) + \epsilon_p e^{\tau_p \tilde{z}_i - \frac{\tau_p^2}{2} - b_i \tau_p^2}}\bigg]. \tag{5.12}$$

Inserting (5.10)-(5.12) into (5.9) gives

$$IV = \int \Pi_{i=1}^k \Big(e^{\sqrt{n}(\omega_i,\tilde{\beta})\tilde{z}_i - \frac{n}{2}(\omega_i,\tilde{\beta})^2} \big[a_i + b_i + (e^{\tau_p^2/n} - 1)\frac{b_i^2}{a_i+b_i} \big] \Big) dF(\tilde{\beta})$$

$$= \int \Pi_{i=1}^k \Big[1 + (e^{\frac{\tau_p^2}{n}} - 1)\epsilon_p^2 e^{\tau_p^2} E\big[\frac{e^{(2+d_i)\tau_p z_i - (2+d_i)^2\tau_p^2/2}}{1-\epsilon_p + \epsilon_p e^{\tau_p \tilde{z}_i - \frac{\tau_p^2}{2} - b_i\tau_p^2}} \big] \Big] dF(\tilde{\beta}). \tag{5.13}$$

Write $\frac{\tau_p^2}{n}\epsilon_p^2 e^{\tau_p^2} E\big[e^{(2\tau_p+d_i)z_i - (2\tau_p+d_i)^2/2}/[(1-\epsilon_p) + \epsilon_p e^{\tau_p\tilde{z}_i - \frac{\tau_p^2}{2} - d_i\tau_p}]\big] = A_i + B_i$, where

$$A_i = \Big(\frac{\tau_p^2}{n}\epsilon_p^2 e^{\tau_p^2}\Big) E\Big[e^{(2\tau_p+b_i)z_i - (2\tau_p+b_i)^2/2} 1_{\{z \le t_p + b_i\}} \Big] = \Big(\frac{\tau_p^2}{n}\epsilon_p^2 e^{\tau_p^2}\Big)\Phi(t_p - 2\tau_p),$$

$$B_i = \Big(\frac{\tau_p^2}{n}\epsilon_p\Big) E\Big[e^{(\tau_p+b_i)z_i - (\tau_p+b_i)^2/2} 1_{\{z > t_p + b_i\}} \Big] = \Big(\frac{\tau_p^2}{n}\epsilon_p\Big)\bar{\Phi}(t_p - \tau_p),$$

and $t_p = [(r+\beta)/(2r)]\tau_p$. First, by Mills' ratio [41], $A_i \le L_p p^{-2\beta+2r-\theta}$. Second, for B_i, noting that $t_p/\tau_p > 1$ in the range of interest, so $B_i \le L_p p^{-(\beta+r)^2/(4r)-\theta}$. By our assumptions, there is a constant $c_0 = c_0(\beta, r, \theta) > 0$ such that $\min\{2\beta - 2r + \theta, \frac{(\beta+r)^2}{4r} + \theta\} \ge 1 + c_0$. Combining these gives

$$\Big(\frac{\tau_p^2}{n}\epsilon_p^2 e^{\tau_p^2}\Big) E\Big[\frac{e^{(2\tau_p+d_i)z_i - (2\tau_p+d_i)^2/2}}{1-\epsilon_p + \epsilon_p e^{\tau_p\tilde{z}_i - \frac{\tau_p^2}{2} - d_i\tau_p}} \Big] \le L_p p^{-(1+c_0)}. \tag{5.14}$$

Inserting (5.14) into (5.13), $IV \le 1 + p^{-c_0}$. Inserting this into (5.8) gives the claim. □

5.3. *Proof of Lemma 1.1.* We define $R_0, R_1, \ldots, R_M$ recursively as follows: (a) Let $R_0 = \emptyset$. (b). Given $R_0, \ldots, R_{m-1}$, let $R_m \subset \{1, 2, \ldots, p\} \setminus (R_0 \cup \ldots \cup R_{m-1})$ be the subset the size of which is as large as possible and satisfies that $\Omega(k, \ell) = 0$ for any two different indices $k \in R_m$ and $\ell \in R_m$ (if there are more than one such subsets, pick any one). The process is repeated until no index is left. Clearly, the constructed $R_1, R_2, \ldots, R_M$ satisfy the second claim, and all remains to show is that $M \le CK_p \log(p)$.

For $m = 0, 1, \ldots, M$, let $s_m = |R_m|$. The key to the proof is that for all $0 \le m \le M$,

$$s_{m+1} \ge \max\{1, \frac{1}{K_p + 1}(p - s_0 - s_1 - \ldots - s_m)\}. \tag{5.15}$$

Since the proofs are similar, we only show the case $m = 0$. Let $R_1 = \{i_1, i_2, \ldots, i_{s_1}\}$, and $D_k = \{1 \le i \le p : \ i \ne i_k, \Omega(i_k, i) \ne 0\}, 1 \le k \le s_1$.

By the assumption of K_p-sparse, $|D_k| \leq K_p$, and so $\big|R_1 \cup \big(\cup_{1\leq k\leq s_1} D_k\big)\big| \leq |R_1| + \sum_{1\leq k\leq s_1}\big|D_k\big| \leq s_1(K_p+1)$. If (5.15) does not hold, then $s_1(K_p+1) < p$, and there is an index $j^* \notin R_1 \cup \big(\cup_{1\leq k\leq s_1} D_k\big)$. Let $R_1^* = \{j^*\} \cup R_1$. It is seen that $\Omega(j,k) = 0$ for any $j,k \in R_1^*$ and $j \neq k$, which contradicts with the definition of R_1. This shows that $s_1(K_p+1) \geq p$ and (5.15) follows.

Next, let $m_0 \leq M$ be the largest indices such that $s_1 + \ldots + s_{m_0} \leq p - K_p - 1$. We claim that for all $1 \leq m \leq m_0$,

$$(5.16)\quad (s_1+s_2+\ldots+s_m) \geq \frac{p}{K_p+1}\sum_{j=0}^{m-1}(1-\frac{1}{K_p+1})^j \equiv p[1-(\frac{K_p}{K_p+1})^m].$$

It suffices to show the first inequality. We show this by mathematical induction. First, by (5.15), this is true for $m = 1$. Second, if this holds for $m-1$, then $(s_1 + s_2 + \ldots s_{m-1}) \geq p\sum_{j=1}^{m-1}\frac{K_p^{m-1}}{(K_p+1)^j}$. At the same time, by (5.15), $s_1 + \ldots + s_m \geq (s_1 + \ldots + s_{m-1}) + \frac{1}{K_p+1}(p - [s_1 + \ldots + s_{m-1}]) = \frac{p}{K_p+1} + \frac{K_p}{K_p+1}(s_1 + \ldots + s_{m-1})$. Combining these with basic algebra, the inequality holds for m and the claim follows.

By (5.16), $1 - (K_p/(K_p+1))^{m_0} \leq (p - K_p - 1)/p$. Therefore, $m_0 \leq \log(\frac{p}{K_p+1})/\log(1+\frac{1}{K_p}) \leq C(K_p+1)\log(p)$. Also, by the way the sets are constructed, $M \leq m_0 + K_p + 1$. Combining these gives the claim. □

5.4. *Proof of Lemmas 2.1-2.2.* Before we prove these two lemmas, we need some preparations. Recall that $D_j = \{k : 1 \leq k \leq p, \Omega(j,k) \neq 0\}$ for $1 \leq j \leq p$. Introduce events $A_{0j} = \{\mu(k) = 0, \forall k \in D_j\}$, $A_{1j} = \{\mu(k) \neq 0 \text{ for exactly one } k \in D_j\}$, and $A_{2j} = \{\mu(k) \neq 0 \text{ for some } k \in D_j, k \neq j\}$. Let $\tilde{\mu} = \Omega\mu$. It is seen that

- Over the event A_{0j}, $\tilde{\mu}(j) = 0$.
- Over the event $A_{1j} \cap \{\mu(j) \neq 0\}$, $\sqrt{n_p}\tilde{\mu}(j) = \sqrt{n_p}\mu(j) = \tau_p$.
- Over the event $A_{1j} \cap \{\mu(j) = 0\}$, $\sqrt{n_p}|\tilde{\mu}(j)| \leq a\tau_p$.

Let $h_0(t) = h_0(t,\epsilon_p,\tau_p,\Omega) = p^{-1}\sum_{j=1}^p P(|\tilde{Z}(j)| \geq t; A_{0j})$, $h_1^+(t) = h_1^+(t,\epsilon_p,\tau_p,\Omega) = p^{-1}\sum_{j=1}^p P(\tilde{Z}(j) \geq t; A_{1j}\cap\{\mu(j)\neq 0\})$, $h_1^-(t) = h_1^-(t,\epsilon_p,\tau_p,\Omega) = p^{-1}\sum_{j=1}^p P(\tilde{Z}(j) \leq -t; A_{1j} \cap \{\mu(j) \neq 0\})$, and $g_2(t) = \frac{\sqrt{n_p}}{p\tau_p}\sum_{j=1}^p E[\tilde{\mu}(j)\mathrm{sgn}(\tilde{Z}(j))\cdot 1\{|\tilde{Z}(j)| \geq t\}|A_{2j}]P(A_{2j})$. Further, recall that $g_1(t) = \frac{1}{p}\sum_{j=1}^p P(|\tilde{Z}(j)| \geq t, A_{2j})$. By definitions, it follows that

$$(5.17)\quad \widetilde{F}(t) = h_0(t) + h_1^+(t) + h_1^-(t) + g_1(t),\ m_p(t) = n_p^{-1/2}p\tau_p\big(h_1^+(t) - h_1^-(t) + g_2(t)\big).$$

Lemma 5.2 below summarizes some basic properties of these quantities, the proof of which is elementary so we omit it.

LEMMA 5.2. *For any $t > 0$, we have (a) $(1 - K_p\epsilon_p)\bar{\Psi}(t) \le h_0(t) \le \bar{\Psi}(t)$, (b) $(1 - K_p\epsilon_p)\epsilon_p\bar{\Phi}(t - \tau_p) \le h_1^+(t) \le \epsilon_p\bar{\Phi}(t - \tau_p)$, $(1 - K_p\epsilon_p)\epsilon_p\bar{\Phi}(t + \tau_p) < h_1^-(t) \le \epsilon_p\bar{\Phi}(t + \tau_p)$, (c) $0 < g_1(t) \le K_p\epsilon_p\bar{\Psi}_{a\tau_p}(t) + (K_p\epsilon_p)^2\bar{\Psi}_{(1+a)\tau_p}(t) + C(K_p\epsilon_p)^3$, (d) $0 \le g_2(t) \le K_p g_1(t)$, and (e) $(1 - K_p\epsilon_p)(\bar{\Psi}(t) + \epsilon_p\bar{\Psi}_{\tau_p}(t)) \le \widetilde{F}(t)$.*

Next, the following lemma is proved in the appendix.

LEMMA 5.3. *Fix $a \in (0, 1)$ and $\tau > 0$. Let (X, Y) be a bivariate normal distribution with mean vector $(0, \tau)'$, variance one and correlation ρ. Then there is a constant $C = C(a) > 0$ such that for all $\rho \in [-a, a]$, $P(|X| \ge t||Y| \ge t) \le C(1 + t)\exp\big(-\frac{(1-a)t^2}{2(1+a)}\big)$.*

By Lemma 5.3, we have the following lemma which is proved in Section 5.9.

LEMMA 5.4. *For any $t > 0$, we can write $v_p(t) = p\big(\widetilde{F}(t) + rem(t)\big)$, where the reminder term $rem(t)/\widetilde{F}(t)$ can be bounded from above by*

(5.18)
$$\begin{cases} L_p p^{-\min\{r, \frac{\beta-r}{2}, (1-a)(\beta-ar)\}} + L_p(1+t)\exp\big(-\frac{(1-a)t^2}{2(1+a)}\big), & r < \beta \text{ and } t \le \tau_p + \tilde{s}_p, \\ K_p, & r \ge \beta \text{ or } t > \tau_p + \tilde{s}_p, \end{cases}$$

where $\tilde{s}_p = \sqrt{\max\{2(\beta - r), (\beta + r)\}\log p}$. Moreover, when $r < \beta$ and $t \le \tau_p + \tilde{s}_p$, we have $v_p(t)/(p\widetilde{F}(t)) \ge 1 - o(1)$. In addition, if the smallest eigenvalue of Ω is bounded from below by $b > 0$, then $v_p(t)/[p\widetilde{F}(t)] \ge b$.

Recall that in (2.17) and (2.8), we defined $W_0(t)$ and its proxy $\widetilde{W_0}(t)$, respectively. Define $a(t) = \sqrt{p}(W_0(t))^{-1}[h_1^+(t) + h_1^-(t) + g_1(t)](v_p(t))^{-1/2}$ and $S_1(t) = (v_p(t))^{-1/2}[\sqrt{p}(g_2(t) - g_1(t) - 2h_1^-(t))]$. Then $\widetilde{Sep}(t, \epsilon_p, \tau_p, \Omega) = 2\tau_p\sqrt{p/n_p}[a(t)W_0(t) + S_1(t)]$. The following two lemmas are proved in Section 5.7 and 5.8.

LEMMA 5.5. *Fix $(\beta, r) \in (0, 1)^2$ and $\Omega \in \mathcal{M}_p^*(a, K_p)$. Then*

(5.19)
$$\sup_{\{0<t\le\tau_p+\tilde{s}_p\}} |S_1(t)| \le L_p\big(p^{-3\beta/2} + p^{-(\beta+r)}\big) + L_p p^{-c_0(\beta,r,a)} \sup_{\{0<t<\infty\}} \widetilde{W_0}(t),$$

where $c_0(\beta, r, a)$ is defined in (2.12) and $\tilde{s}_p$ is defined in Lemma 5.4. If in addition $\Omega \in \widetilde{\mathcal{M}}_p^*(a, b, K_p)$, then the above inequality holds with the left hand side replaced with $\sup_{\{t>0\}} |S_1(t)|$.

Also, if $r < \beta$ *and* $t \le \tau_p + \tilde{s}_p$, *then* $|a(t)-1| \le L_p p^{-\min\{r, \frac{\beta-r}{2}, (1-a)(\beta-ar)\}} + L_p(1+t)\exp\big(-\frac{(1-a)t^2}{2(1+a)}\big)$; *and if in addition* $\Omega \in \widetilde{\mathcal{M}}_p^*(a, b, K_p)$, *then* $K_p^{-1/2} \lesssim a(t) \lesssim b^{-1/2}$.

LEMMA 5.6. *Fix* $(r, \beta) \in (0,1)^2$. *Then*

$$\sup_{\{t>0\}} |W_0(t) - \widetilde{W}_0(t)| \le L_p p^{-3\beta/2} + 2 \sup_{\{t>0\}} \frac{K_p \epsilon_p \bar{\Psi}_{a\tau_p}(t)}{\sqrt{\bar{\Psi}(t) + K_p \epsilon_p \bar{\Psi}_{a\tau_p}(t)}} = L_p p^{-3\beta/2} + L_p p^{-c_0(\beta,r,a)} \sup_{\{t>0\}} \widetilde{W}_0(t), \tag{5.20}$$

where $c_0(\beta, r, a)$ *is defined in (2.12).*

We now prove Lemma 2.1 and Lemma 2.2 separately.

5.5. *Proof of Lemma 2.1.* Write for short $\widetilde{Sep}(t) = \widetilde{Sep}(t, \epsilon_p, \tau_p, \Omega)$. We consider the two cases 1) $t > \tau_p + \tilde{s}_p$ and 2) $t \le \tau_p + \tilde{s}_p$ separately, where $\tilde{s}_p$ is as in Lemma 5.4.

First consider case 1). We will show that (1a) $\widetilde{Sep}(t) \le L_p p^{\frac{1-\theta}{2} - \max\{\beta - \frac{1}{2}r, \frac{3\beta+r}{4}\}}$ and (1b) $\widetilde{W}_0(t) \le L_p p^{-\max\{\beta-\frac{1}{2}r, \frac{3\beta+r}{4}\}}$. Then combining (1a) and (1b) completes the proof of the lemma in case 1). We now proceed to prove (1a) and (1b). The result (1b) follows immediately from the definition of $\widetilde{W}_0(t)$ and the inequalities $\widetilde{W}_0(t) \le \sqrt{\epsilon_p \bar{\Psi}_{\tau_p}(t)} \le L_p p^{-\max\{4\beta-2r, 3\beta+r\}/4}$. It remains to prove (1a). Let η be a $p \times 1$ vector such that $\eta(j) = 1\{(\Omega \hat{\mu}_t^{\tilde{Z}})(j) \ne 0\}$, $1 \le j \le p$. Also, for any $p \times 1$ vectors x and y, let $x \circ y$ be the $p \times 1$ vector such that $(x \circ y)(j) = x(j)y(j)$, $1 \le j \le p$. By definition, it is seen that $m_p(t) = E[\tilde{M}_p(t)] = E[(\hat{\mu}_t^{\tilde{Z}})'\Omega\mu] = E[(\hat{\mu}_t^{\tilde{Z}})'\Omega(\mu \circ \eta)]$. Using Cauchy-Schwartz inequality, $m_p(t) \le \big(E[(\hat{\mu}_t^{\tilde{Z}})'\Omega\hat{\mu}_t^{\tilde{Z}}]\big)^{1/2}\big(E[(\mu\circ\eta)'\Omega(\mu\circ\eta)]\big)^{1/2}$. Recalling that $v_p(t) = E[\tilde{V}_p(t)] = E[(\hat{\mu}_t^{\tilde{Z}})'\Omega\hat{\mu}_t^{\tilde{Z}}]$, it follows that

$$|\widetilde{Sep}(t)| = 2m_p(t)(v_p(t))^{-1/2} \le 2\big(E[(\mu\circ\eta)'\Omega(\mu\circ\eta)]\big)^{1/2}. \tag{5.21}$$

Since the largest eigenvalue of Ω is no greater than K_p, the last term above $\le 2K_p^{1/2}(E\|\mu\circ\eta\|^2)^{1/2}$ and so $|\widetilde{Sep}(t)| \le 2K_p^{1/2}(E\|\mu\circ\eta\|^2)^{1/2}$. It remains to study $E\|\mu\circ\eta\|^2$. By definition,

$$E\|\mu\circ\eta\|^2 = \sum_{i=1}^p \frac{\tau_p^2}{n_p} P(\mu(i) \ne 0, (\Omega\hat{\mu}_t^{\tilde{Z}})(i) \ne 0) \le \frac{\tau_p^2}{n_p}\sum_{i=1}^p \sum_{j\in D_i} P(\mu(i)\ne 0, \hat{\mu}_t^{\tilde{Z}}(j) \ne 0)$$

$$= \frac{\tau_p^2}{n_p}\sum_{i=1}^p \sum_{j\in D_i} P(\mu(i)\ne 0, |\tilde{Z}(j)| \ge t) \le L_p p^{1-\theta}(\epsilon_p \bar{\Psi}_{\tau_p}(t) + \epsilon_p \bar{\Psi}_{a\tau_p}(t) + CK_p\epsilon_p^2).$$

Since we consider the range $t > \tau_p + \tilde{s}_p$, the above expectation can be bounded as $E\|\mu \circ \eta\|^2 \leq L_p p^{1-\theta-\max\{4\beta-2r,3\beta+r\}/2}$. Inserting this into (5.21) we complete the proof of (1a).

Now we consider the case 2). Recall that $\widetilde{Sep}(t) = 2\tau_p\sqrt{p/n_p}[a(t)W_0(t) + S_1(t)]$. Noting that $n_p = p^\theta$, the key is to show

$$\sup_{\{0<t\leq\tau_p+\tilde{s}_p\}} \Big|(2\tau_p)^{-1}p^{(\theta-1)/2}\widetilde{Sep}(t) - W_0(t)\Big| \leq L_p p^{-3\beta/2} + L_p p^{-\beta-r} + L_p\Big(p^{-\min\{r,\frac{\beta-r}{2},(1-a)(\beta-ar)\}} + p^{-c_0(\beta,r,a)} + p^{-\tilde{c}_1(\beta,r,a)}\Big)\sup_{\{t>0\}}\widetilde{W}_0(t)\Big). \tag{5.22}$$

In fact, once this is proved, the claim follows by using Lemma 5.6.

We now show (5.22). By Lemma 5.5,

$$\begin{aligned}&\sup_{\{0<t\leq\tau_p+\tilde{s}_p\}}|p^{(\theta-1)/2}(2\tau_p)^{-1}\widetilde{Sep}(t) - W_0(t)| \\ &\leq \sup_{\{0<t\leq\tau_p+\tilde{s}_p\}}|a(t)-1|W_0(t) + \sup_{\{0<t\leq\tau_p+\tilde{s}_p\}}|S_1(t)|.\end{aligned} \tag{5.23}$$

The second term on the right was studied in Lemma 5.5 inequality (5.19). We now study the first term on the right. By lemma 5.5,

$$\sup_{\{0<t\leq\tau_p+\tilde{s}_p\}}|a(t)-1|W_0(t) \leq \sup_{\{t\geq 0\}} I_1(t) + \sup_{\{t\geq 0\}} I_2(t), \tag{5.24}$$

where $I_1(t) = L_p\big(p^{-\min\{r,\frac{\beta-r}{2},(1-a)(\beta-ar)\}} + C(1+t)\exp(-\frac{1-a}{2(1+a)}t^2)\big)\widetilde{W}_0(t)$, and $I_2(t) = L_p|W_0(t) - \widetilde{W}_0(t)|$.

Consider $I_2(t)$ first. By Lemma 5.6 and Lemma 5.5,

$$\sup_{\{t\geq 0\}} I_2(t) \leq L_p\Big(p^{-3\beta/2} + p^{-c_0(\beta,r,a)}\sup_{\{0<t<\infty\}}\widetilde{W}_0(t)\Big). \tag{5.25}$$

Consider $I_1(t)$ next. Write $I_1(t) = I_{1a}(t) + I_{1b}(t)$, where $I_{1a}(t) = L_p \cdot p^{-\min\{r,\frac{\beta-r}{2},(1-a)(\beta-ar)\}}\widetilde{W}_0(t)$ and $I_{1b}(t) = L_p(1+t)\exp(-\frac{1-a}{2(1+a)}t^2)\widetilde{W}_0(t)$. We first study $I_{1b}(t)$. By definitions and elementary algebra,

$$\sup_{\{0<t<\infty\}}\{(1+t)\exp(-\frac{(1-a)}{2(1+a)}t^2)\widetilde{W}_0(t)\} = L_p p^{-\tilde{c}_1(\beta,r,a)}\sup_{\{0<t<\infty\}}\widetilde{W}_0(t),$$

where $\tilde{c}_1(\beta,r,a)$ is defined in (2.12). Combining these results and comparing terms yields

$$\sup_{t>0} I_1(t) \leq L_p\Big(p^{-\min\{r,\frac{\beta-r}{2},(1-a)(\beta-ar)\}} + p^{-\tilde{c}_1(\beta,r,a)}\Big)\sup_{\{0<t<\infty\}}\widetilde{W}_0(t). \tag{5.26}$$

Combing (5.26) and (5.25) with (5.24) yields

$$\sup_{\{0<t\le\tau_p+\tilde{s}_p\}}|a(t)-1|W_0(t)\le L_p p^{-3\beta/2}$$
$$+\Big(p^{-\min\{r,\frac{\beta-r}{2},(1-a)(\beta-ar)\}}+p^{-c_0(\beta,r,a)}+p^{-\tilde{c}_1(\beta,r,a)}\Big)\sup_{\{0<t<\infty\}}\widetilde{W}_0(t).$$

Inserting this and (5.19) into (5.23) completes the proof of the lemma when $t\le\tau_p+\tilde{s}_p$. □

5.6. *Proof of Lemma 2.2.* First, we consider (a)-(b). By Lemma 5.5, $(2\tau_p)^{-1}\sqrt{n_p/p}\widetilde{Sep}(t)\le b^{-1/2}W_0(t)+S_1(t)$, where $W_0(t)$ is defined in (2.17), and $S_1(t)$ is as in Lemma 5.5. The key is to prove that there is a constant $d_0>0$ such that for any fixed t satisfying either $0\le t\le\sqrt{2\beta\log p}-d_0\log\log p/\sqrt{\log p}$ or $t>\tau_p+2\sqrt{\log(K_p\log p)}$,

$$W_0(t)\lesssim\ \tfrac{2\sqrt{b\epsilon_p}}{3K_p},\qquad S_1(t)\lesssim\tfrac{\sqrt{b\epsilon_p}}{3K_p(\log p)}. \tag{5.27}$$

In fact, once these are proved, then

$$\widetilde{Sep}(t)\le 2\tau_p p^{(1-\theta)/2}[b^{-1/2}W_0(t)+S_1(t)]\lesssim\frac{5}{3}\tau_p K_p^{-1}p^{(1-\theta-\beta)/2}, \tag{5.28}$$

and parts (a)-(b) of the lemma follow.

We now show (5.27). Recall that by the proof of Lemmas 5.5-5.6,

$$|S_1(t)|\le L_p(p^{-3\beta/2}+p^{-\beta-r})+\frac{CK_p\epsilon_p\bar{\Psi}_{a\tau_p}(t)}{\sqrt{\bar{\Psi}(t)+K_p\epsilon_p\bar{\Psi}_{a\tau_p}(t)}}, \tag{5.29}$$

$$0<W_0(t)-\widetilde{W}_0(t)\le L_p p^{-3\beta/2}+\frac{CK_p\epsilon_p\bar{\Psi}_{a\tau_p}(t)}{\sqrt{\bar{\Psi}(t)+K_p\epsilon_p\bar{\Psi}_{a\tau_p}(t)}}; \tag{5.30}$$

note that the last terms in the above two inequalities are the same. We now consider the case $t\le\sqrt{2\beta\log p}-d_0\log\log p/\sqrt{\log p}$ and the case $t>\tau_p+2\sqrt{\log(K_p\log p)}$ separately.

In the first case, by Mills's ratio [41], with the constant $d_0>0$ being appropriately chosen, $\bar{\Psi}(t)+K_p\epsilon_p\bar{\Psi}_{a\tau_p}(t)\ge 9C^2b^{-1}K_p^4(\log p)^2\epsilon_p$ and $\bar{\Psi}(t)+\epsilon_p\bar{\Psi}_{\tau_p}(t)\ge 9b^{-1}K_p^2\epsilon_p$. As a result,

$$\frac{CK_p\epsilon_p\bar{\Psi}_{a\tau_p}(t)}{\sqrt{\bar{\Psi}(t)+K_p\epsilon_p\bar{\Psi}_{a\tau_p}(t)}}\le\frac{\sqrt{b\epsilon_p}}{3K_p\log p},\quad \widetilde{W}_0(t)=\frac{\epsilon_p\bar{\Psi}_{\tau_p}(t)}{\sqrt{\bar{\Psi}(t)+\epsilon_p\bar{\Psi}_{\tau_p}(t)}}\le\frac{\sqrt{b\epsilon_p}}{3K_p}.$$

Inserting these into (5.29) and (5.30), the claim follows by noting that $\epsilon_p = p^{-\beta}$.

Consider the second case. In this case, $\epsilon_p\bar{\Psi}_{a\tau_p}(t) = o(\epsilon_p p^{-(1-a)^2 r})$. Thus,

$$\frac{K_p\epsilon_p\bar{\Psi}_{a\tau_p}(t)}{\sqrt{\bar{\Psi}(t) + K_p\epsilon_p\bar{\Psi}_{a\tau_p}(t)}} \le \sqrt{K_p\epsilon_p\bar{\Psi}_{a\tau_p}(t)} = o(K_p^{-1}(\log p)^{-1}\sqrt{\epsilon_p}),$$

and

$$\widetilde{W}_0(t) = \frac{\epsilon_p\bar{\Psi}_{\tau_p}(t)}{\sqrt{\bar{\Psi}(t) + \epsilon_p\bar{\Psi}_{\tau_p}(t)}} \le \sqrt{\epsilon_p\bar{\Psi}_{\tau_p}(t)} \lesssim \sqrt{b\epsilon_p}/(3K_p).$$

Inserting these into (5.29) and (5.30) proves (5.27), the claim follows by similar reasons.

Next, consider (c). Write for short $s_p = \sqrt{2\beta\log p} - d_0\log\log p/\sqrt{\log p}$. Since the eigenvalue of Ω is bounded from above by K_p, by definition we have $v_p(t) \le K_p p\widetilde{F}(t)$. Thus, $\widetilde{Sep}(t) = 2m_p(t)/\sqrt{v_p(t)} \ge 2K_p^{-1/2}m_p(t)/\sqrt{p\widetilde{F}(t)}$. By definitions in (5.17) and Lemma 5.2 we can further obtain that

$$\widetilde{Sep}(t) \ge \frac{2\tau_p p^{\frac{1-\theta}{2}}(h_1^+(t) - h_1^-(t))}{\sqrt{K_p\widetilde{F}(t)}} \ge \frac{2\tau_p p^{\frac{1-\theta}{2}}[(1-K_p\epsilon_p)\epsilon_p\bar{\Phi}(t-\tau_p) - \epsilon_p\bar{\Phi}(t+\tau_p)]}{\sqrt{K_p\big(\bar{\Psi}(t) + \epsilon_p\bar{\Psi}_{\tau_p}(t) + K_p\epsilon_p\bar{\Psi}_{a\tau_p}(t) + C(K_p\epsilon_p)^2\big)}}.$$

When $s_p \le t \le \tau_p$, the numerator above $\sim 2\tau_p p^{\frac{1-\theta}{2}-\beta}$, and the denominator above $\le K_p p^{-\frac{\beta}{2}}$. Thus, $\widetilde{Sep}(t) \ge 2\tau_p K_p^{-1}p^{(1-\theta-\beta)/2}$. On the other hand, recall that $\sup_{t>0}\widetilde{W}_0(t) = L_p p^{-\beta/2}$ when $r \ge \beta$, which together with Lemmas 5.5-5.6 ensures $\sup_{t>0}W_0(t) \le L_p p^{-\beta/2}$ and $\sup_{t>0}S_1(t) \le L_p p^{-\beta/2}$. Since $(2\tau_p)^{-1}\sqrt{n_p/p}\widetilde{Sep}(t) \le b^{-1/2}W_0(t) + S_1(t)$, combining these entails $\widetilde{Sep}(t) \le L_p p^{(1-\theta-\beta)/2}$. This completes the proof of part (c). □

5.7. *Proof of Lemma 5.6.* Recall that $W_0(t) = \frac{\epsilon_p\bar{\Psi}_{\tau_p}(t)+g_1(t)}{\sqrt{\bar{\Psi}(t)+\epsilon_p\bar{\Psi}_{\tau_p}(t)+g_1(t)}}$, where $g_1(t)$ is as in Lemma 5.2. We will compare $W_0(t)$ with $\widetilde{W}_0(t)$ defined in (2.8). On one hand, since $(A+x)/\sqrt{B+x}$ is an increasing function of x when $0 \le A < B$, it is seen that $W_0(t) \ge \widetilde{W}_0(t)$. On the other hand, writing for short $b(t) = K_p\epsilon_p\bar{\Psi}_{a\tau_p}(t) + (K_p\epsilon_p)^2\bar{\Psi}_{(1+a)\tau_p}(t)$, it follows from Lemma 5.2(c)

that

(5.31)

$$W_0(t) \le \frac{\epsilon_p \bar\Psi_{\tau_p}(t) + b(t) + CK_p^3\epsilon_p^3}{\sqrt{\bar\Psi(t) + \epsilon_p\bar\Psi_{\tau_p}(t) + b(t) + CK_p^3\epsilon_p^3}}$$
$$\le \widetilde{W}_0(t) + CK_p^{3/2}p^{-3\beta/2} + \frac{K_p\epsilon_p\bar\Psi_{a\tau_p}(t)}{\sqrt{\bar\Psi(t) + K_p\epsilon_p\bar\Psi_{a\tau_p}(t)}} + \frac{K_p^2\epsilon_p^2\bar\Psi_{(1+a)\tau_p}(t)}{\sqrt{\epsilon_p\bar\Psi_{\tau_p}(t) + b(t)}}.$$

Combining these and recalling $\epsilon_p = p^{-\beta}$, we have

$$\sup_{0<t<\infty} |W_0(t) - \widetilde{W}_0(t)| \le L_p p^{-3\beta/2} + I + II,$$

where

$$I = \sup_{0<t<\infty} \frac{K_p\epsilon_p\bar\Psi_{a\tau_p}(t)}{\sqrt{\bar\Psi(t) + K_p\epsilon_p\bar\Psi_{a\tau_p}(t)}}, \qquad II = \sup_{0<t<\infty} \frac{K_p^2\epsilon_p^2\bar\Psi_{(1+a)\tau_p}(t)}{\sqrt{\epsilon_p\bar\Psi_{\tau_p}(t) + b(t)}}.$$

To show the first inequality of claim, it is sufficient to show

(5.32)

$$II \le L_p p^{-3\beta/2} + L_p p^{-\beta/2} \sup_{0<t<\infty} \frac{K_p\epsilon_p\bar\Psi_{a\tau_p}(t)}{\sqrt{\bar\Psi(t) + K_p\epsilon_p\bar\Psi_{a\tau_p}(t)}} \equiv L_p p^{-3\beta/2} + L_p p^{-\beta/2}\cdot I.$$

Towards this end, we write $II \le IIa + IIb$, where IIa and IIb are the supremum of $K_p^2\epsilon_p^2\bar\Psi_{(1+a)\tau_p}(t)/\sqrt{\epsilon_p\bar\Psi_{\tau_p}(t) + b(t)}$ over the intervals $0 < t < \tau_p$ and $\tau_p \le t < \infty$, respectively. Consider IIa. When $0 \le t \le \tau_p$, $\bar\Psi_{\tau_p}(t) \ge 1/2$, and so $IIa \le K_p^2\epsilon_p^2 \sup_{\{0<t<\tau_p\}} \frac{\bar\Psi_{(1+a)\tau_p}(t)}{\sqrt{\epsilon_p\bar\Psi_{\tau_p}(t)}} \le L_p\epsilon_p^{3/2}$. Consider IIb. By definitions and change-of-variable, and recalling $\epsilon_p = p^{-\beta}$,

$$IIb \le \sup_{\{\tau_p\le t<\infty\}} \frac{K_p^2\epsilon_p^2\bar\Psi_{(1+a)\tau_p}(t)}{\sqrt{\epsilon_p\bar\Psi_{\tau_p}(t) + K_p^2\epsilon_p^2\bar\Psi_{(1+a)\tau_p}(t)}} = \sup_{\{0\le t<\infty\}} \frac{K_p^2\epsilon_p^{3/2}\bar\Psi_{a\tau_p}(t)}{\sqrt{\bar\Psi(t) + K_p^2\epsilon_p\bar\Psi_{a\tau_p}(t)}}$$
$$\le L_p\epsilon_p^{1/2}\cdot I = L_p p^{-\beta/2}\cdot I.$$

Combining these proves (5.32). Consequently, the first inequality of the claim follows.

To show the second inequality in the claim, we use similar calculations as in [16] and get

$$\sup_{\{0\le t<\infty\}} \{\widetilde{W}_0(t)\} = L_p p^{-\delta(r,\beta)},\ I = L_p p^{-\delta(a^2r,\beta)} \equiv L_p p^{-c_0(\beta,r,a)} \sup_{0<t<\infty} \widetilde{W}_0(t),$$

where we have used $c_0(r,\beta,a) = \delta(r,\beta) - \delta(a^2r,\beta)$ as in (2.10). □

5.8. *Proof of Lemma 5.5.* Consider the first claim. By Lemma 5.2 (part (d)), $|g_2(t)| \le K_p g_1(t)$. So by definitions,

$$|S_1(t)| \le (K_p+1)\frac{\sqrt{p}g_1(t)}{\sqrt{v_p(t)}} + \frac{2\sqrt{p}h_1^-(t)}{\sqrt{v_p(t)}} \equiv (K_p+1)B_0(t) + B_1(t). \tag{5.33}$$

Consider $B_0(t)$ first. Rewrite $B_0(t) = [g_1(t)/\sqrt{\widetilde{F}(t)}]\sqrt{p\widetilde{F}(t)/v_p(t)}$. Note that when $r < \beta$ and $t \le \tau_p + \tilde{s}_p$, $p\widetilde{F}(t)/v_p(t) \lesssim 1$, and when $r \ge \beta$ and $\Omega \in \widetilde{M}^*(a,b,K_p)$, by the last claim of Lemma 5.4, $p\widetilde{F}(t)/v_p(t) \le b^{-1}$. This says that $p\widetilde{F}(t)/v_p(t) \le C$ for some generic constant $C > 0$ and so $B_0(t) \le Cg_1(t)/\sqrt{\widetilde{F}(t)}$. At the same time, by definitions and Lemma 5.2, $\widetilde{F}(t) = h_0(t) + h_1^+(t) + h_1^-(t) + g_1(t) \ge (1 - K_p\epsilon_p)[\bar{\Psi}(t) + \epsilon_p\bar{\Psi}_{\tau_p}(t)] + g_1(t)$, so we have

$$B_0(t) \le Cg_1(t)/\sqrt{\bar{\Psi}(t) + \epsilon_p\bar{\Psi}_{\tau_p}(t) + g_1(t)}.$$

Finally, using Lemma 5.2 and noting that $x/\sqrt{A+x}$ is an increasing function in $x \in (0,\infty)$ for any number $A > 0$, we obtain

$$B_0(t) \le \frac{C\big(K_p\epsilon_p\bar{\Psi}_{a\tau_p}(t) + (K_p\epsilon_p)^2\bar{\Psi}_{(1+a)\tau_p}(t) + (K_p\epsilon_p)^3\big)}{\sqrt{\bar{\Psi}(t) + \epsilon_p\bar{\Psi}_{\tau_p}(t) + K_p\epsilon_p\bar{\Psi}_{a\tau_p}(t) + (K_p\epsilon_p)^2\bar{\Psi}_{(1+a)\tau_p}(t) + (K_p\epsilon_p)^3}}.$$

where the right hand side $\le I + II + C(K_p\epsilon_p)^{3/2}$, with

$$I = \frac{CK_p\epsilon_p\bar{\Psi}_{a\tau_p}(t)}{\sqrt{\bar{\Psi}(t) + K_p\epsilon_p\bar{\Psi}_{a\tau_p}(t)}},\ II = \frac{C(K_p\epsilon_p)^2\bar{\Psi}_{(1+a)\tau_p}(t)}{\sqrt{\epsilon_p\bar{\Psi}_{\tau_p}(t) + K_p\epsilon_p\bar{\Psi}_{a\tau_p}(t) + (K_p\epsilon_p)^2\bar{\Psi}_{(1+a)\tau_p}(t)}}.$$

The above two terms have been considered in Lemma 5.6 (see the last two terms of (5.31)). Using the results over there we can show that

$$\sup_{\{0<t\le\tilde{s}_p\}} B_0(t) \le L_p p^{-3\beta/2} + L_p p^{-c_0(\beta,r,a)} \sup_{\{0<t<\infty\}} \widetilde{W}_0(t). \tag{5.34}$$

Next we consider $B_1(t)$. Write $B_1(t) = 2\cdot[(p\widetilde{F}(t)/v_p(t))^{1/2}]\cdot[h_1^-(t)(\widetilde{F}(t))^{-1/2}]$. We have just proved $p\widetilde{F}(t)/v_p(t) \le C$ when $r \ge \beta$ or $0 < t \le \tau_p + \tilde{s}_p$ with $C > 0$ some generic constant. At the same time, using (5.17) and parts (a)-(b) of Lemma 5.2, first, $h_1^-(t) \le \epsilon_p\bar{\Phi}(t+\tau_p)$, and second, $\widetilde{F}(t) \ge h_0(t) + h_1^+(t) + h_1^-(t) \ge (1 - K_p\epsilon_p)[\bar{\Psi}(t) + \epsilon_p\bar{\Psi}_{\tau_p}(t)]$. Combining these gives $h_1^-(t)(\widetilde{F}(t))^{-1/2} \le C\epsilon_p\bar{\Phi}(t+\tau_p)/\sqrt{\bar{\Psi}(t) + \epsilon_p\bar{\Psi}_{\tau_p}(t)}$. It follows that $B_1(t) \le$

$C\epsilon_p\bar{\Phi}(t+\tau_p)/\sqrt{\bar{\Psi}(t)+\epsilon_p\bar{\Psi}_{\tau_p}(t)}$. This together with direct calculations yields

$$\sup_{0<t\le\tilde{s}_p} B_1(t) \le C\epsilon_p \sup_{0<t<\infty} \frac{\bar{\Phi}(t+\tau_p)}{\sqrt{\bar{\Psi}(t)+\epsilon_p\bar{\Psi}_{\tau_p}(t)}} = Cp^{-(\beta+r)}. \tag{5.35}$$

Inserting (5.34) and (5.35) into (5.33) completes the proof.

Consider the last two claims. Write $a(t) = A_1 \cdot A_2 \cdot A_3$, where

$$A_1 = \frac{h_1^+(t)+h_1^-(t)+g_1(t)}{\epsilon_p\bar{\Psi}_{\tau_p}(t)+g_1(t)}, \quad A_2 = \Big(\frac{\bar{\Psi}(t)+\epsilon_p\bar{\Psi}_{\tau_p}(t)+g_1(t)}{\widetilde{F}(t)}\Big)^{1/2},$$

and $A_3 = \big(p\widetilde{F}(t)/v_p(t)\big)^{1/2}$. First, by Lemma 5.2 (part (b)), $\epsilon_p(1-K_p\epsilon_p)\bar{\Psi}_{\tau_p}(t) \le h_1^+(t)+h_1^-(t) \le \epsilon_p\bar{\Psi}_{\tau_p}(t)$ and thus $1-K_p\epsilon_p \le A_1 \le 1$. Second, similarly, by Lemma 5.2, $1 \le A_2 \le (1-K_p\epsilon_p)^{-1/2}$. Since by basis algebra, $|AB-1| \le |A-1|+|B-1|+|A-1||B-1|$ for any numbers A and B, we have $|a(t)-1| \le CK_p\epsilon_p(1+|A_3-1|)+|A_3-1|$. Now, by Lemma 5.4, $|A_3-1| \le L_p\Big(p^{-\min\{r,\frac{\beta-r}{2},(1-a)(\beta-ar)\}} + (1+t)\exp\big(-\frac{(1-a)t^2}{2(1+a)}\big)\Big)$ when $r<\beta$ and $0<t\le\tau_p+\tilde{s}_p$, and $K_p^{-1/2} \le A_3(t) \le b^{-1/2}$ when $\Omega \in \widetilde{\mathcal{M}}_p^*(a,b,K_p)$, and so the claim follows. □

5.9. *Proof of Lemma 5.4.* The last claim follows trivially from the assumption on the minimum eigenvalue of Ω. And in the case of $r \ge \beta$, by definition of $v_p(t)$ and noting that the maximum eigenvalue of Ω is bounded by K_p, we obtain that $v_p(t) \le K_p p\widetilde{F}(t)$. So we only need to prove the first claim in the case of $r<\beta$ and the second claim.

Consider the first claim. Let $D_i = \{j : \Omega(i,j) \ne 0\}$ and $\tilde{D}_i = D_i \setminus \{i\}$. Write $h(t) = \tilde{h}_0(t)+\tilde{h}_1(t)$, where $h(t) = p^{-1}\sum_{i=1}^p \sum_{j\in\tilde{D}_i} P(|\tilde{Z}(i)| \ge t, |\tilde{Z}(j)| \ge t)$, $\tilde{h}_0(t) = p^{-1}\sum_{i,j\in\tilde{D}_i} P(|\tilde{Z}(i)| \ge t, |\tilde{Z}(j)| \ge t, \tilde{\mu}_i = 0 \text{ or } \tilde{\mu}_j = 0)$, $\tilde{h}_1(t) = p^{-1}\sum_{i,j\in\tilde{D}_i} P(|\tilde{Z}(i)| \ge t, |\tilde{Z}(j)| \ge t, \tilde{\mu}_i \ne 0 \text{ and } \tilde{\mu}_j \ne 0)$. By definitions, it is seen that

$$v_p(t) = p(\widetilde{F}(t)+rem(t)), \text{ where } |rem(t)| \le h(t) = \tilde{h}_0(t)+\tilde{h}_1(t). \tag{5.36}$$

To show the claim, it is sufficient to show that the ratio $[\tilde{h}_0(t)+\tilde{h}_1(t)]/\widetilde{F}(t)$ does not exceed the right hand side of (5.18).

First, consider $\tilde{h}_0(t)$. If at least one of $\tilde{Z}(i)$ and $\tilde{Z}(j)$ has mean 0, by Lemma 5.3 and definitions, $P(|\tilde{Z}(i)| \ge t, |\tilde{Z}(j)| \le t, \tilde{\mu}_j = 0 \text{ or } \tilde{\mu}_j = 0) \le CK_p(1+t)\exp(-\frac{(1-a)t^2}{2(1+a)})\big(P(|\tilde{Z}(i)| \ge t)+P(|\tilde{Z}(j)| \ge t)\big)$. Since $\tilde{D}_i$ has at

most K_p components, it follows from the definition of $\widetilde{F}(t)$ that

(5.37)
$$\begin{aligned}\tilde{h}_0(t) &\le CK_p(1+t)\exp(-\frac{(1-a)t^2}{2(1+a)})p^{-1}\sum_{i,j\in\tilde{D}_i}\big(P(|\tilde{Z}(i)|\ge t)+P(|\tilde{Z}(j)|\ge t)\big)\\ &\le CK_p^2(1+t)\exp\big(-\frac{(1-a)t^2}{2(1+a)}\big)\widetilde{F}(t).\end{aligned}$$

Next, consider $\tilde{h}_1(t)$. Define events $A_{1,ij}=\{\mu(k)\neq 0$ for some $k\in D_i\setminus D_j\}$, $A_{2,ij}=\{\mu(k)\neq 0$ for exactly one k, which is in $D_i\cap D_j\}$, and $A_{3,ij}=\{\mu(k)\neq 0$ for two or more k, all of which are in $D_i\cap D_j\}$. It is seen that

$$\begin{aligned}\tilde{h}_1(t) &= p^{-1}\sum\nolimits_{i,j\in\tilde{D}_i}P(|\tilde{Z}(i)|\ge t,|\tilde{Z}(j)|\ge t,\tilde{\mu}(i)\neq 0,\tilde{\mu}(j)\neq 0)\\ &= \tilde{h}_{1,1}(t)+\tilde{h}_{1,2}(t)+\tilde{h}_{1,3}(t),\end{aligned}$$

where $\tilde{h}_{1,1}(t)=p^{-1}\sum_{i,j\in\tilde{D}_i}P(|\tilde{Z}(i)|\ge t,|\tilde{Z}(j)|\ge t,A_{1,ij}\cup A_{1,ji})$, $\tilde{h}_{1,2}(t)=p^{-1}\sum_{i,j\in\tilde{D}_i}P(|\tilde{Z}(i)|\ge t,|\tilde{Z}(j)|\ge t,A_{2,ij})$, and $\tilde{h}_{1,3}(t)=p^{-1}\sum_{i,j\in\tilde{D}_i}P(|\tilde{Z}(i)|\ge t,|\tilde{Z}(j)|\ge t,A_{3,ij})$.

We first consider $\tilde{h}_{1,1}(t)$. Note that

$$\begin{aligned}&P(|\tilde{Z}(i)|\ge t,|\tilde{Z}(j)|\ge t,A_{1,ij}\cup A_{1,ji})\\ &\le P(|\tilde{Z}(i)|\ge t,|\tilde{Z}(j)|\ge t,A_{1,ji})+P(|\tilde{Z}(i)|\ge t,|\tilde{Z}(j)|\ge t,A_{1,ij})\\ &\le P(|\tilde{Z}(i)|\ge t,A_{1,ji})+P(|\tilde{Z}(j)|\ge t,A_{1,ij})\le K_p\epsilon_p[P(|\tilde{Z}(i)|\ge t)+P(|\tilde{Z}(j)|\ge t)].\end{aligned}$$

Thus, $\tilde{h}_{1,1}(t)\le 2\epsilon_pK_p^2p^{-1}\sum_{i=1}^pP(|\tilde{Z}(i)|\ge t)=L_p\epsilon_p\widetilde{F}(t)$.

Now we consider $\tilde{h}_{1,2}(t)$. For any $(i,j)\in A_{2,ij}$, we use $(\tilde{Z}^*(i),\tilde{Z}^*(j))$ to denote the demeaned pair of $(\tilde{Z}(i),\tilde{Z}(j))$. By definition there exists a k such that $\sqrt{n_p}\mu(k)=\tau_p$, $\tilde{\mu}(i)=\Omega(i,k)\mu(k)$ and $\tilde{\mu}(j)=\Omega(j,k)\mu(k)$. Thus, $|\sqrt{n_p}\tilde{\mu}(i)|\le a\tau_p$ or $|\sqrt{n_p}\tilde{\mu}(j)|\le a\tau_p$ and

$$P(|\tilde{Z}(i)|\ge t,|\tilde{Z}(j)|\ge t,A_{2,ij})\le K_p\epsilon_pP(|\tilde{Z}^*(i)|\ge t-a\tau_p)=K_p\epsilon_p\bar{\Psi}_{a\tau_p}(t).$$

Then $\tilde{h}_{1,2}(t)\le K_p^2\epsilon_p\bar{\Psi}_{a\tau_p}(t)$. Direct calculations yield

(5.38)
$$\frac{\tilde{h}_{1,2}(t)}{\bar{\Psi}(t)+\epsilon_p\bar{\Psi}_{\tau_p}(t)}\le\frac{K_p^2\epsilon_p\bar{\Psi}_{a\tau_p}(t)}{\bar{\Psi}(t)+\epsilon_p\bar{\Psi}_{\tau_p}(t)}\le L_pp^{-(1-a)(\beta-ar)},\ \text{for all } t\le\tau_p+\tilde{s}_p.$$

By Lemma 5.4 $\widetilde{F}(t)\gtrsim\bar{\Psi}(t)+\epsilon_p\bar{\Psi}_{\tau_p}(t)$, it follows that $\tilde{h}_{1,2}(t)\le L_pp^{-(1-a)(\beta-ar)}\widetilde{F}(t)$.

Now, consider $\tilde{h}_{1,3}(t)$. Observe that $\tilde{h}_{1,3}(t) \leq p^{-1}\sum_{i,j\in\tilde{D}_i} P(A_{3,ij}) \leq K_p(K_p\epsilon_p)^2$. By Lemma 5.2,

$$\frac{\tilde{h}_{1,3}(t)}{\widetilde{F}(t)} \leq \frac{1}{1-K_p\epsilon_p}\frac{K_p(K_p\epsilon_p)^2)}{\bar{\Psi}(t)+\epsilon_p\bar{\Psi}_{\tau_p}(t)} \leq \frac{CK_p^3\epsilon_p^2}{\bar{\Psi}(t)+\epsilon_p\bar{\Psi}_{\tau_p}(t)}.$$

When $r<\beta$ and $t\leq\tau_p+\tilde{s}_p$, we have $\bar{\Psi}(t)+\epsilon_p\bar{\Psi}_{\tau_p}(t) \geq L_p p^{-\max\{4\beta-2r,3\beta+r\}/2}$, and thus $CK_p^3\epsilon_p^2/[\bar{\Psi}(t)+\epsilon_p\bar{\Psi}_{\tau_p}(t)] \leq L_p(p^{-(\beta-r)/2}+p^{-r})$. When $t>\tau_p+\tilde{s}_p$, by the definition of $v_p(t)$ and recalling that the largest eigenvalue of Ω is bounded by K_p, we have $v_p(t)\leq K_p p\widetilde{F}(t)$. Combining these together and noting that $\widetilde{F}(t)\gtrsim\bar{\Psi}(t)+\epsilon_p\bar{\Psi}_{\tau_p}(t)$, we obtain $\tilde{h}_{1,3}(t)/\widetilde{F}(t)\leq K_p$ if $t>\tau_p+\tilde{s}_p$, and $\tilde{h}_{1,3}(t)/\widetilde{F}(t)\leq L_p(p^{-(\beta-r)/2}+p^{-r})$ if $t\leq\tau_p+\tilde{s}_p$.

Combining the bounds on $\tilde{h}_{1,1}(t),\tilde{h}_{1,2}(t)$ and $\tilde{h}_{1,3}(t)$ entails that when $r<\beta$, $\tilde{h}_1(t)/\widetilde{F}(t)\leq p^{-(\beta-r)/2}+p^{-r}+p^{-(1-a)(\beta-ar)}$ if $t\leq\tau_p+\tilde{s}_p$ and $\tilde{h}_1(t)/\widetilde{F}(t)\leq K_p$ if $t>\tau_p+\tilde{s}_p$. These together with (5.36) and (5.37) completes the proof of the first claim when $r<\beta$.

Next, we consider the second claim. The goal is to show that $v_p(t)/(p\widetilde{F}(t))\gtrsim 1$, assuming $r<\beta$ and $t\leq\tau_p+\tilde{s}_p$. We consider the cases (a) $d_3\log\log p\leq t\leq\tau_p+\tilde{s}_p$ and (b) $t<d_3\log\log p$ separately, where $d_3>0$ is a large constant.

In Case (a), using (5.37), it is seen that $|rem(t)|/\widetilde{F}(t)|=o(1)$, uniformly for all $d_3\log\log p\leq t\leq\tau_p+\tilde{s}_p$. Using (5.36), $|v_p(t)/(p\widetilde{F}(t))-1|=o(1)$ and the claim follows.

In Case (b), recall that $v_p(t)=E[(\hat{\mu}_t^{\tilde{Z}})'\Omega\mu_t^{\tilde{Z}}]$, where $\hat{\mu}_t^{\tilde{Z}}(j)=\mathrm{sgn}(\tilde{Z}(j))1\{|\tilde{Z}(j)|\geq t\}$ and $\tilde{Z}=\Omega Z$. Write $\tilde{Z}=\sqrt{n_p}\tilde{\mu}+W$, where $\tilde{\mu}=\Omega\mu$ and $W\sim N(0,\Omega)$. Let $\hat{\mu}_t$ be the counterpart of $\hat{\mu}_t^{\tilde{Z}}$ defined by $\hat{\mu}_t(j)=\mathrm{sgn}(W(j))1\{|W(j)|\geq t\}$. We claim (b1) $E[(\hat{\mu}_t^{\tilde{Z}})'\Omega\mu_t^{\tilde{Z}}]=E[(\hat{\mu}_t)'\Omega\hat{\mu}_t]+O(L_pp^{1-\beta/2})$ and (b2) $E[(\hat{\mu}_t)'\Omega\hat{\mu}_t]\geq p\widetilde{F}(t)$. The claim follows by combining (b1) and (b2) and noting that $p\widetilde{F}(t)\geq L_pp(1-K_p\epsilon_p)$ when $t\leq d_3\log\log p$.

Consider (b1). Let $S=\{1\leq i\leq p:\hat{\mu}_t^{\tilde{Z}}(i)\neq\hat{\mu}_t(i)\}$. Note that for all $p\times 1$ vectors ξ and η, by Schwartz inequality and that the spectral norm of $\Omega\leq K_p$, $|(\xi+\eta)'\Omega(\xi+\eta)-\eta'\Omega\eta|\leq\xi'\Omega\xi+2[(\xi'\Omega\xi)\cdot(\eta'\Omega\eta)]^{1/2}\leq L_p[\|\xi\|^2+\|\xi\|\|\eta\|]$. Applying this with $\eta=\hat{\mu}_t$, $\xi=\hat{\mu}_t^{\tilde{Z}}-\hat{\mu}_t$, and noting that each coordinate of $\hat{\mu}_t^{\tilde{Z}}-\hat{\mu}_t$ has magnitude no greater than 2, we claim that $|E[(\hat{\mu}_t^{\tilde{Z}})'\Omega\mu_t^{\tilde{Z}}]-E[(\hat{\mu}_t)'\Omega\hat{\mu}_t]|\leq L_pE[|S|+\sqrt{p|S|}]\leq L_pE[\sqrt{p|S|}]$. Note that for any $i\in S$, we must have $\tilde{\mu}(i)\neq 0$. Therefore, by definitions, $|S|\leq\sum_{i=1}^p 1\{(\Omega\mu)(i)\neq 0\}\leq\sum_{i=1}^p\sum_{j:\Omega(i,j)\neq 0}1\{\mu(j)\neq 0\}\leq K_p\sum_{i=1}^p 1\{\mu(i)\neq 0\}$, where we have used the assumption that Ω is K_p-sparse. Note that $\sum_{i=1}^p 1\{\mu(i)\neq 0\}\sim\mathrm{Binomial}(p,\epsilon_p)$, where $\epsilon_p=p^{-\beta}$, so $E[\sqrt{p|S|}]\sim p^{1-\beta/2}$.

Combining these gives (b1).

Consider (b2). Denoting $B = E[\hat{\mu}_t\hat{\mu}_t']$, we have $E[(\hat{\mu}_t)'\Omega\hat{\mu}_t] = E[\Omega\hat{\mu}_t\hat{\mu}_t'] = tr(\Omega B)$. We claim that for any $i \neq j$ such that $\Omega(i,j) \neq 0$, $B(i,j)$ has the same sign as that of $\Omega(i,j)$. To see the point, write $B(i,j) = E[\mathrm{sgn}(\tilde{Z}(i))\mathrm{sgn}(\tilde{Z}(j))\cdot 1\{|\tilde{Z}(i| > t, |\tilde{Z}(j)| > t\}$. By symmetry and basic statistics, $B(i,j) = 2[P(\tilde{Z}(j) > t, \tilde{Z}(j) > t|\Omega(i,j)) - P(\tilde{Z}(i) > t, \tilde{Z}(j) > t| - \Omega(i,j))]$, where for any $\rho \in (-1,1)$, $P(\tilde{Z}(i) > t, \tilde{Z}(j) > t|\rho)$ is evaluated at the law that $\mathrm{corr}(\tilde{Z}(i), \tilde{Z}(j)) = \rho$. The claim follows by noting that for any $\rho > 0$, $P(\tilde{Z}(j) > t, \tilde{Z}(j) > t|\rho) > P(\tilde{Z}(i) > t)P(\tilde{Z}(j) > t) > P(\tilde{Z}(i) > t, \tilde{Z}(j) > t| - \rho)$. As a result, $tr(\Omega B) \geq tr(B) \equiv p\widetilde{F}(t)$, where we have used the fact that the diagonals of Ω are ones. This proves (b2). □

5.10. *Proof of Lemmas 2.3-2.4.* Write for short $W(t) = p^{-1/2}HC(t,\widetilde{F})$. Recalling $W_0(t) = [\epsilon_p\bar{\Psi}_{\tau_p}(t) + g_1(t)]/\sqrt{\bar{\Psi}(t) + \epsilon_p\bar{\Psi}_{\tau_p}(t) + g_1(t)}$ as defined in (2.17), where $g_1(t)$ is as in Lemma 5.2, we let $a_1(t) = \big(W_0(t)\big)^{-1}[\widetilde{F}(t) - h_0(t)]\cdot(\widetilde{F}(t)(1-\widetilde{F}(t))^{-1/2}$, and $W_1(t) = [\bar{\Psi}(t) - h_0(t)]\cdot(\widetilde{F}(t)(1-\widetilde{F}(t))^{-1/2}$, where $h_0(t)$ is as in Lemma 5.2. By these notations, $W(t) = a_1(t)W_0(t) - W_1(t)$. The following Lemma is proved in Section 5.11.

LEMMA 5.7. *Fix a sufficiently large p. There is a universal constant $C > 0$ such that for all $\Omega \in \mathcal{M}_p^*(a, K_p)$,*

(5.39)
$$0 < W_1(t) \leq CK_p\epsilon_p\bar{\Psi}(t)/\sqrt{\bar{\Psi}(t) + \epsilon_p\bar{\Psi}_{\tau_p}(t)}, \text{ for all } t \geq \bar{\Psi}^{-1}(1/2)$$

(5.40)
$$1 - CK_p\epsilon_p \leq a_1(t) \leq (1 + CK_p\epsilon_p)(1 - \bar{\Psi}(t) - K_p\epsilon_p)^{-1/2}, \text{ for all } t \geq 0.$$

Consider Lemma 2.3. Using Lemma 5.7, $|a_1(t) - 1| \leq C(K_p\epsilon_p + \bar{\Psi}(t))$ for all $t \geq 0$. Recalling $W(t) = a_1(t)W_0(t) - W_1(t)$, we have

(5.41)
$$\begin{aligned}\sup_{\{t \geq \bar{\Psi}^{-1}(\frac{1}{2})\}} |W(t) - W_0(t)| &\leq \sup_{\{t\geq 0\}}\{|a_1(t) - 1|W_0(t)\} + \sup_{\{t \geq \bar{\Psi}^{-1}(\frac{1}{2})\}} W_1(t)\\ &\leq L_p(I + II + III),\end{aligned}$$

where $I = K_p\epsilon_p \sup_{\{t\geq 0\}}\{W_0(t)\}$, $II = \sup_{\{t\geq 0\}}\{\bar{\Psi}(t)W_0(t)\}$, and $III = \sup_{\{t \geq \bar{\Psi}^{-1}(\frac{1}{2})\}}\{W_1(t)\}$.

First, consider I. By basic algebra and Lemma 5.6,

$$I \leq L_p\epsilon_p[\sup_{\{t\geq 0\}} \widetilde{W}_0(t) + \sup_{t\geq 0}|W_0(t) - \widetilde{W}_0(t)|] \leq L_p p^{-\beta}[p^{-3\beta/2} + \sup_{\{t\geq 0\}}\{\widetilde{W}_0(t)\}].$$

Next, consider II. Write

$$II \le \sup_{\{t\ge 0\}}[\bar\Psi(t)\widetilde{W}_0(t)] + \sup_{\{t\ge 0\}}[\bar\Psi(t)|W_0(t)-\widetilde{W}_0(t)|] \equiv IIa + IIb. \tag{5.42}$$

On one hand, elementary calculus shows that $IIa \le p^{-\beta}$. On the other hand, by similar argument as in the proof of Lemma 5.6, $IIb \le L_p(p^{-\beta}+p^{-\frac{a^2r}{3}-\beta}+p^{-3\beta/2})$. Combining these, $II \le L_p(p^{-\beta}+p^{-\frac{a^2r}{3}-\beta}+p^{-3\beta/2})$. Last, consider III. By (5.39) and direct calculations,

$$III \le CK_p\epsilon_p \sup_{\{t\ge 0\}}\{\bar\Psi(t)/\sqrt{\bar\Psi(t)+\epsilon_p\bar\Psi(t-\tau_p)}\} \le L_p p^{-\beta}.$$

Inserting these into (5.41) gives the claim.

Next, we show Lemma 2.4. The first claim has already been proved in Lemma 5.6. So we only need to prove claims (a)–(c) in the case of $r \ge \beta$.

First consider claims (a) and (b) in Lemma 2.4. Comparing Lemma 5.6 and the desired claim, it is sufficient to verify that

$$W_0(t) \le p^{-\beta/2}/\sqrt{2}, \text{ if } t \le \sqrt{2\beta\log p} - \Delta_1 \text{ or } t > \tau_p, \tag{5.43}$$

where $\Delta_1 = d_0(\log\log p)/\sqrt{\log p}$ is as defined in the statement of Lemma 2.4. Once this is proved, recalling that $W(t) = a_1(t)W_0(t) - W_1(t)$ and we have just proved $\sup_{t\ge\bar\Psi^{-1}(1/2)}\{\bar\Psi(t)W_0(t)\} \le L_p p^{-\beta}$, then by lemma 5.7 we have

$$W(t) \le a(t)W_0(t) \lesssim (1 + C\bar\Psi(t) + CK_p\epsilon_p)W_0(t) \lesssim p^{-\beta/2}/\sqrt{2}.$$

We now proceed to prove (5.43). By the proof of Lemma 5.6 (inequality (5.31)), we have

$$0 \le W_0(t) - \widetilde{W}_0(t) \le L_p p^{-\beta} + K_p\epsilon_p\bar\Psi_{a\tau_p}(t)/\sqrt{\bar\Psi(t)+K_p\epsilon_p\bar\Psi_{a\tau_p}(t)}, \tag{5.44}$$

where we have noted that the last term in (5.31) is bounded by $K_p\epsilon_p\sqrt{\bar\Psi_{(1+a)\tau_p}(t)} \le L_p p^{-\beta}$. First consider the case when $t \le \sqrt{2\beta\log p} - \Delta_1$. By Mills's ratio, for appropriately chosen d_0 in $\Delta_1 = d_0(\log\log p)/\sqrt{\log p}$, we have $\bar\Psi(t) + K_p\epsilon_p\bar\Psi_{a\tau_p}(t) \ge 8K_p^2\epsilon_p$, and $\bar\Psi(t) + \epsilon_p\bar\Psi_{a\tau_p}(t) \ge 8\epsilon_p$. As a result,

$$\frac{K_p\epsilon_p\bar\Psi_{a\tau_p}(t)}{\sqrt{\bar\Psi(t)+K_p\epsilon_p\bar\Psi_{a\tau_p}(t)}} \le \sqrt{2\epsilon_p}/4, \qquad \widetilde{W}_0(t) \le \frac{\epsilon_p\bar\Psi_{\tau_p}(t)}{\sqrt{\bar\Psi(t)+\epsilon_p\bar\Psi_{\tau_p}(t)}} \le \sqrt{2\epsilon_p}/4.$$

Inserting these into (5.44), we complete the proof of (5.43) when $t \le \sqrt{2\beta\log p} - \Delta_1$. Now we consider the case of $t > \tau_p$. Since $\epsilon_p \bar{\Psi}_{a\tau_p}(t) = o(\epsilon_p p^{-(1-a)^2 r})$, it follows that

$$\frac{K_p\epsilon_p\bar{\Psi}_{a\tau_p}(t)}{\sqrt{\bar{\Psi}(t) + K_p\epsilon_p\bar{\Psi}_{a\tau_p}(t)}} \le \sqrt{K_p\epsilon_p\bar{\Psi}_{a\tau_p}(t)} = o(p^{-\beta/2})$$

and

$$\widetilde{W}_0(t) \le \frac{\epsilon_p\bar{\Psi}_{\tau_p}(t)}{\sqrt{\bar{\Psi}(t) + \epsilon_p\bar{\Psi}_{\tau_p}(t)}} \le \sqrt{\epsilon_p\bar{\Psi}_{\tau_p}(t)} \le \sqrt{\epsilon_p/2}.$$

Inserting these into (5.44) proves (5.43) when $t > \tau_p$.

Finally we prove part (c). Write for short $s_p = \sqrt{2\beta\log p} - \Delta_1$. By (5.40) and recalling that we have just proved $\sup_{t>0} W_1(t) \le L_p p^{-\beta}$, we obtain that $W(t) = a_1(t)W_0(t) - W_1(t) \ge (1 - K_p\epsilon_p)W_0(t) - \sup_{t>0} W_1(t) \ge (1 - CK_p\epsilon_p)W_0(t) - L_p p^{-\beta}$. Further recall that in Lemma 5.6, we have shown that $W_0(t) \ge \widetilde{W}_0(t)$ for all $t \ge 0$. Thus, $W(t) \ge (1 - CK_p\epsilon_p)\widetilde{W}_0(t) - L_p p^{-\beta}$. Taking $t_p^* = \frac{\beta+r}{2r}\tau_p$, it is seen that for sufficiently large p, $s_p \le t_p^* \le \tau_p$. Therefore, $\sup_{\{s_p\le t\le\tau_p\}} W(t) \ge (1 - CK_p\epsilon_p)\sup_{\{s_p\le t\le\tau_p\}} \widetilde{W}_0(t) \ge (1 - CK_p\epsilon_p)\widetilde{W}_0(t_p^*)$, and the first inequality of part c) follows from $\widetilde{W}_0(t_p^*) \sim p^{-\beta/2}$. On the other hand, by Lemma 5.6 and recall $r \ge \beta$, we have $\sup_{t>0} W_0(t) \le L_p \sup_{t>0} \widetilde{W}_0(t) \sim L_p p^{-\beta/2}$. Further, by (5.40) and the expression $W(t) = a_1(t)W_0(t) - W_1(t)$, we have $\sup_{s_p\le t\le\tau_p} W(t) \le \sup_{s_p\le t\le\tau_p}\{a_1(t)W_0(t)\} \le C\sup_{s_p\le t\le\tau_p} W_0(t) \sim L_p p^{-\beta/2}$. Thus, the second inequality in the claim follows. □

5.11. *Proof of Lemma 5.7.* Let $h_0(t)$, $h_1^{\pm}(t)$ and $g_1(t)$ be as in Lemma 5.4. Consider the first claim. By Lemma 5.2 parts (a) and (e), we have

$$0 \le \bar{\Psi}(t) - h_0(t) \le K_p\epsilon_p\bar{\Psi}(t), \qquad \widetilde{F}(t) \ge (1 - K_p\epsilon_p)[\bar{\Psi}(t) + \epsilon_p\bar{\Psi}_{\tau_p}(t)]. \tag{5.45}$$

At the same time, note that $\widetilde{F}(t) \le \bar{\Psi}(t) + K_p\epsilon_p$. Combining these ensures that

$$1 \le (1 - \widetilde{F}(t))^{-1/2} \le [1 - \bar{\Psi}(t) - K_p\epsilon_p]^{-1/2}. \tag{5.46}$$

Inserting (5.45) and (5.46) into the definition of $W_1(t)$ gives

$$0 \le W_1(t) \le \frac{K_p\epsilon_p\bar{\Psi}(t)}{\sqrt{(1 - \bar{\Psi}(t) - K_p\epsilon_p)(1 - K_p\epsilon_p)[\bar{\Psi}(t) + \epsilon_p\bar{\Psi}_{\tau_p}(t)]}}.$$

Thus the first claim follows by noting $(1-\bar{\Psi}(t)-K_p\epsilon_p) \geq 1/2 - K_p\epsilon_p$ for all $t \geq \bar{\Psi}^{-1}(\frac{1}{2})$.

Consider the second claim. Recall that $\widetilde{F}(t) = h_0(t) + h_1^+(t) + h_1^{-1} + g_1(t)$. By definitions,

$$a_1(t) = (1-\widetilde{F}(t))^{-1/2} \cdot I \cdot II, \tag{5.47}$$

where $I = [h_1^+(t) + h_1^-(t) + g_1(t)]/[\epsilon_p \bar{\Psi}_{\tau_p}(t) + g_1(t)]$, and

$$II = \sqrt{\bar{\Psi}(t) + \epsilon_p \bar{\Psi}(t-\tau_p) + g_1(t)}/\sqrt{h_0(t) + h_1^+(t) + g_1(t)}.$$

By (a) and (b) in Lemma 5.2, we have

$$(1-K_p\epsilon_p) \leq I \leq 1, \qquad 1 \leq II \leq (1-K_p\epsilon_p)^{-1/2}. \tag{5.48}$$

Inserting (5.46) and (5.48) into (5.47), we obtain that there is a universal constant $C > 0$ such that (5.40) holds. □

5.12. *Proof of Theorems 2.1-2.2.* The following lemma is proved in Section 5.13.

LEMMA 5.8. *Fix $(\beta, r) \in (0,1)^2$ and a sufficiently large p. When t ranges in $(0,\infty)$, $\widetilde{W}_0(t)$ first strictly increases and reaches the maximum at $t = t_p^{**} \sim \min\{2, \frac{r+\beta}{2r}\}\tau_p\,(\equiv t_p^*)$, and then strictly decreasing. Additionally, if $r < \beta$, then there are positive constants $c_4 = c_4(\beta, r)$ and $c_5 = c_5(\beta, r)$ such that for all $|t - t_p^{**}| \leq c_4\tau_p^{-1}$, $\widetilde{W}_0''(t) \leq -2c_5\widetilde{W}_0(t)$.*

Denote by $W(t) = p^{-1/2}HC(t, \widetilde{F})$. By the first claim in Lemma 2.3 and Lemma 5.6, and noting that $\beta > c_0(\beta, r, a)$, we obtain

$$\sup_{\{t\geq 0\}} |W(t) - \widetilde{W}_0(t)| \leq L_p[p^{-\beta} + p^{-c_0(\beta,r,a)} \sup_{\{t\geq 0\}} \widetilde{W}_0(t)]. \tag{5.49}$$

First, we show Theorem 2.1, where we assume $r < \beta$. Once the first claim is proved, the second claim follows by combining Taylor expansion with Lemmas 2.3, 2.4, and 5.8, so we only show the first claim. The idea is to prove T_{HC} and T_{ideal} are both close to t_p^{**}, then they are close to each other.

We first prove that T_{HC} and t_p^{**} are close. We will show that (i) $W(t_p^{**} + u) - W(t_p^{**}) < 0$ for all $|u| \leq c_4/\tau_p$, and (ii) $W(t) - W(t_p^{**}) < 0$ for all $|u| > c_4/\tau_p$. Then combining these proves

$$|T_{HC}(\widetilde{F}) - t_p^{**}| \leq p^{-c_1}, \tag{5.50}$$

with $c_1 = c_1(\beta, r, a) > 0$ some constant to be specified later.

We now prove the first case (i). Recall that t_p^{**} is the maximizer of $\widetilde{W}_0(t)$ and $\widetilde{W}_0(t_p^{**}) = L_p p^{-\delta(\beta,r)}$, where $\delta(\beta, r)$ is as in (2.10). Thus, $\widetilde{W}_0'(t_p^{**}) = 0$. By Taylor expansion, $\widetilde{W}_0(t_p^{**} + u) - \widetilde{W}_0(t_p^{**}) = \frac{u^2}{2}\widetilde{W}_0''(\tilde{t}_p)$, where $\tilde{t}_p$ lies between t_p^{**} and $t_p^{**} + u$. Next, by Lemma 5.8, for $|u| \le \frac{c_4}{\tau_p}$ we can further write $\widetilde{W}_0(t_p^{**} + u) - \widetilde{W}_0(t_p^{**}) \le -c_5 u^2 \widetilde{W}_0(\tilde{t}_p) = -c_5 u^2 \widetilde{W}_0(t_p^{**}) - c_5 u^2(\widetilde{W}_0(\tilde{t}_p) - \widetilde{W}_0(t_p^{**})) \le -c_5 u^2 \widetilde{W}_0(t_p^{**}) - c_5 u^2(\widetilde{W}_0(t_p + u) - \widetilde{W}_0(t_p^{**}))$, where the last step is because of $\widetilde{W}_0(t_p^{**} + u) \le \widetilde{W}_0(\tilde{t}_p)$. Thus, the inequality can be further written as $\widetilde{W}_0(t_p^{**} + u) - \widetilde{W}_0(t_p^{**}) \le -c_5 u^2 \widetilde{W}_0(t_p^{**})/(1 + c_5 u^2)$. Then by (5.49) we obtain that

(5.51)
$$
\begin{aligned}
&W(t_p^{**} + u) - W(t_p^{**}) = \big(W(t_p^{**} + u) - \widetilde{W}_0(t_p^{**} + u)\big) - \big(W(t_p^{**}) - \widetilde{W}_0(t_p^{**})\big) \\
&+ \big(\widetilde{W}_0(t_p^{**} + u) - \widetilde{W}_0(t_p^{**})\big) \le L_p(p^{-\beta} + p^{-c_0(\beta,r,a)}\widetilde{W}_0(t_p^{**})) + \big(\widetilde{W}_0(t_p^{**} + u) - \widetilde{W}_0(t_p^{**})\big) \\
&\le L_p p^{-\beta} + \big(L_p p^{-c_0(\beta,r,a)} - c_5 u^2/(1 + c_5 u^2)\big)\widetilde{W}_0(t_p^{**})
\end{aligned}
$$

It is easy to check that $p^{-c_0(\beta,r,a)}\widetilde{W}_0(t_p^{**}) \gg L_p p^{-\beta}$ when $\rho_\theta^*(\beta) < r < \beta$. By Lemma 5.8, we obtain that if $|u| \ge p^{-c_1}$ with $c_1 = c_1(\beta, r, a) \in (0, \frac{1}{3}c_0(\beta, r, a))$, then for all $|u| \le c_4/\tau_p$,

$$W(t_p^{**} + u) - W(t_p^{**}) \le -L_p p^{-2c_1(\beta,r,a)}\widetilde{W}_0(t^*)(1 + o(1)) < 0,$$

which completes the proof of case (i). It remains to prove case (ii). Direct calculations yield $\widetilde{W}_0(t_p^{**} \pm c_4/\tau_p) \lesssim e^{-c_5}\widetilde{W}_0(t_p^{**})$, where $c_5 > 0$ is a constant depending on whether $r < \beta/3$ or $r \ge \beta/3$. By Lemma 5.8, $\widetilde{W}_0(t) \le \widetilde{W}_0(t_p^{**} \pm c_4/\tau_p) \lesssim e^{-c_5}\widetilde{W}_0(t_p^{**})$ for all $|t - t_p^{**}| > c_4/\tau_p$. Thus, similar to (5.51) we have $W(t) - W(t_p^{**}) \le L_p(p^{-\beta} + p^{-c_0(\beta,r,a)}\widetilde{W}_0(t_p^{**})) + (\widetilde{W}_0(t) - \widetilde{W}_0(t_p^{**})) \lesssim L_p p^{-\beta} + (e^{-c_5} - 1 + L_p p^{-c_0(\beta,r,a)})\widetilde{W}_0(t_p^{**}) = L_p p^{-\beta} + (e^{-c_5} - 1 + L_p p^{-c_0(\beta,r,a)})p^{-\delta(\beta,r)} < 0$, where the last step is because $\beta > \delta(\beta, r)$. This proves case (ii). Consequently, we have proved (5.50).

Using similar method as above and in view of Lemma 2.1 we can also prove that for appropriately chosen $c_1 > 0$,

$$|T_{ideal}(\epsilon_p, \tau_p, \Omega) - t_p^{**}| \le p^{-c_1}. \tag{5.52}$$

Thus the claim in Theorem 2.1 follows when $r < \beta$.

We now show Theorem 2.2, where we assume $r \ge \beta$. In this range $\widetilde{W}_0(t)$ is maximized at $t_p^{**} = \frac{\beta + r}{2r}\tau_p$ and $\widetilde{W}_0(t_p^{**}) \sim p^{-\frac{\beta}{2}}$. By Lemma 2.4 we see that

the maximizer of $W_0(t)$ is in the range $[\sqrt{2\beta\log p}-\Delta_1,\tau_p)$. By (5.43) and Lemma 2.3 we obtain that if $0\le t<\sqrt{2\beta\log p}-\Delta_1$ or $\tau_p\le t<\infty$,

$$W(t)=W_0(t)+(W(t)-W_0(t))\le\frac{1}{\sqrt{2}}p^{-\beta/2}+L_pp^{-\beta}=\frac{1}{\sqrt{2}}p^{-\beta/2}(1+o(1)),$$

and if $\sqrt{2\beta\log p}-\Delta_1\le t<\tau_p$,

$$W(t)=W_0(t)+(W(t)-W_0(t))\ge p^{-\beta/2}-L_pp^{-\beta}=p^{-\beta/2}(1-o(1)).$$

Thus, the maximizer $T_{HC}(\widetilde{F})$ is in the interval $[\sqrt{2\beta\log p}-\Delta_1,\tau_p)$.

By Lemma 2.2, the maximizer of $\widetilde{Sep}(t,\epsilon_p,\tau_p,\Omega)$ is in the interval $[\sqrt{2\beta\log p}-\Delta_1,\tau_p+\Delta_2)$. Thus, Theorem 2.2 follows immediately from Lemma 2.2. □

5.13. *Proof of Lemma 5.8.* Let $\psi_{\tau_p}(t)=\phi(t-\tau_p)+\phi(t+\tau_p)$ and $\psi(t)=2\phi(t)$. Introduce $m_0(t)=\psi(t)/\bar{\Psi}(t)$, $m_1(t)=\bar{\Psi}_{\tau_p}(t)/\psi_{\tau_p}(t)$, $d(t)=-\psi'_{\tau_p}(t)/\psi_{\tau_p}(t)$, $a(t)=\epsilon_p\psi_{\tau_p}(t)/\psi(t)$, $R(t)=m_1(t)/m_0(t)$, and $g(t)=(1/2)(1+a(t))/(R^{-1}(t)+a(t))$. The following lemma is proved in Section 6.4.

LEMMA 5.9. *Fix a sufficiently large p, $R(t)>1$ and is strictly decreasing for all $t>0$.*

Consider the first claim. By direct calculations and our notations,

$$\widetilde{W}_0'(t)/\widetilde{W}_0(t)=\frac{1}{2}\Big[\frac{\psi(t)+\epsilon_p\psi_{\tau_p}(t)}{\bar{\Psi}(t)+\epsilon_p\bar{\Psi}_{\tau_p}(t)}\Big]-\frac{\psi_{\tau_p}(t)}{\bar{\Psi}_{\tau_p}(t)}\equiv[g(t)-1]/m_1(t). \tag{5.53}$$

To show the claim, it suffices to show that equation $g(t)=1$ has exactly one solution. Recall that $g(t)=(1/2)(1+a(t))/(R^{-1}(t)+a(t))$, where $R(t)>1$ and both $a(t)$ and $R^{-1}(t)$ are strictly increasing in t. It follows from basic calculus that $g(t)$ is strictly decreasing in $(0,\infty)$, and the equation $g(t)=1$ has at most one solution.

The equation also has at least one solution. Note that $g(0)\ge Ce^{\tau_p^2/2}$ which >1 for sufficiently large p, it suffices to show that there is a t such that $g(t)<1$. We show this for the case of $r<\beta/3$ and $r>\beta/3$ separately. In the first case, for all t such that $|t-2\tau_p|\le 4\tau_p^{-1}$, $a(t)$ is algebraically small, and so by Mills' ratio [41], for any fixed b,

$$g(2\tau_p+b\tau_p^{-1})\le\frac{1}{2}\Big[\frac{1}{2}-\frac{3b}{2}\tau_p^{-2}+O(\tau_p^{-4})\Big],$$

and the claim follows. Note that this shows that the solution t_p^{**} of the equation $g(t)=1$ satisfies $|t_p^{**}-2\tau_p|\le 2\tau_p^{-1}$. In the second case, $a(\sqrt{2\log(p)})=$

$L_p p^{1-\beta-(1-\sqrt{r})^2}$, where the the exponent > 0 since $r > \beta/3$ and $r > \rho(\beta)$ (recall that $\rho(\beta)$ is the standard phase function). Therefore, $g(t_0) \sim 1/2$ and the claim follows. This completes the proof of the first claim.

Consider the second claim. We discuss for the case $0 < r < \beta/3$ and $\beta/3 < r < \beta$ separately.

Consider the first case. Recalling that $|t_p^{**} - 2\tau_p| \le 2\tau_p^{-1}$, it is sufficient to show that for all t such that $|t_p - 2\tau_p| \le 4\tau_p^{-1}$, $\widetilde{W}_0''(t)/\widetilde{W}_0(t) \lesssim -1/2$. Introduce $s(t) = [t\psi(t) + d(t)\psi_{\tau_p}(t)] \cdot [\bar{\Psi}(t) + \epsilon_p \bar{\Psi}_{\tau_p}(t)]/[\psi(t) + \epsilon_p \psi_{\tau_p}(t)]^2$. By direct calculations,

$$\widetilde{W}''/\widetilde{W}(t) = I + II - \frac{1}{2}III, \tag{5.54}$$

where

$$I = (g(t)-1)^2/m_1^2(t),\ II = d(t)/m_1(t) - m_1^{-2}(t),\ III = (s(t)-1)g^2(t)m_1^{-2}(t).$$

Consider I first. When $|t - 2\tau_p| \le 4\tau_p^{-1}$, on one hand, by Mills' ratio, $m_1^{-1}(t) \sim (t - \tau_p) \sim \tau_p$. On the other hand, by similar argument, $|g(t) - 1| \le C\tau_p^{-2}$. It follows that $I \le C\tau_p^{-2}$. Consider II next. By Mills' ratio, $m_1^{-1}(t) = (t - \tau_p) + \frac{1}{t-\tau_p} + O(\tau_p^{-3})$. Since $|d(t) - (t - \tau_p)|$ is algebraically small, it follows from basic algebra that $II \sim -1$. Consider III. Note that both the ratio $\epsilon_p \psi_{\tau_p}(t)/\psi(t)$ and the ratio $\epsilon_p \bar{\Psi}_{\tau_p}(t)/\bar{\Psi}_{\tau_p}(t)$ are algebraically small. Combining this with $\bar{\Psi}(t)/\psi(t) = (1/t) - (1/t^3) + O(t^{-5})$ gives

$$s(t) = \frac{t\psi(t)\bar{\Psi}(t)}{(\psi(t))^2} + O(\tau_p^{-3}) = 1 - \frac{1}{t^2} + O(\tau_p^{-3}),$$

Recall that $m_1^{-1}(t) \sim \tau_p$ and $g(t) \sim 1$, it follows that $III \sim -4\tau_p^2/t^2 \sim -1$. Inserting these into (5.54) gives that for all $|t - 2\tau_p| \le 4\tau_p^{-1}$, $\widetilde{W}_0''(t)/\widetilde{W}_0(t) \lesssim -1/2$ and the second claim follows.

Consider the second case, where $r \ge \beta$. For a constant $\eta_0 \in (0, 1)$ to be determined, choose t_0 and $t_p^{\pm}$ such that $a(t_0) = \frac{3r-\beta}{\beta+r}$, and $a(t_p^{\pm}) = (1 \pm \eta_0)a(t_0)$. It is seen that $|t_p^{\pm} - \frac{\beta+r}{2r}\tau_p| \le C\tau_p^{-1}$, and $|t_0 - \frac{\beta+r}{2r}\tau_p| \le C\tau_p^{-1}$. Combining these with definitions and Mills' ratio, for $t_p^- \le t \le t_p^+$, $R^{-1}(t) \sim (t - \tau_p)/t \sim (\beta - r)/(\beta + r)$, and that

$$g(t) \sim \frac{1}{2} \cdot \frac{1 + a(t)}{[(\beta - r)/(\beta + r)] + a(t)}. \tag{5.55}$$

By direct calculations, $g(t_p^-) > 1$ and $g(t_p^+) < 1$. Since $g(t_p^{**}) = 1$, we have $t_p^- < t_p^{**} < t_p^+$.

We now use (5.54) to calculate $\widetilde{W}_0''(t)/\tilde{W}_0(t)$ with. First, recall that $II \sim -1$. Second, by similar argument, $m_1^{-1}(t) \sim (t-\tau_p) \sim (\beta-r)/(2r)\tau_p$. Combining this with (5.55),

$$I = m_1^{-2}(t)[g(t)-1]^2 = (\frac{\beta-r}{2r})^2\tau_p^2\cdot\big([\frac{1}{2}\frac{1+a(t)}{[(\beta-r)/(\beta+r)+a(t)}-1]^2+o(1)\big).$$

Last, by similar argument,

$$\frac{t\psi(t)+\epsilon_p d(t)\psi_{\tau_p}(t)}{\bar{\Psi}(t)+\epsilon_p\bar{\Psi}_{\tau_p}(t)} \sim \frac{(\beta+r)/(\beta-r)+a(t)}{(\beta-r)/(\beta+r)+a(t)}(\frac{\beta-r}{2r})^2\tau_p^2.$$

$$s(t) \sim \frac{[(\beta+r)/(\beta-r)+a(t)]\cdot[(\beta-r)/(\beta+r)+a(t)]}{(1+a(t))^2}.$$

Combining this with (5.55), III equals to $(\frac{\beta-r}{2r})^2\tau_p^2\cdot[\frac{1+a(t)}{(\beta-r)/(\beta+r)+a(t)}]^2$ times

$$\big[\frac{[(\beta+r)/(\beta-r)+a(t)]\cdot[(\beta-r)/(\beta+r)+a(t)]}{(1+a(t))^2}-1+o(1)\big].$$

Inserting these into (5.54) and recalling that $a(t_0)=(3r-\beta)/(\beta+r)$, it follows from basic algebra that $\widetilde{W}''(t_0)/\widetilde{W}(t_0) \lesssim -\frac{3r-\beta}{2(\beta-r)}\cdot(\frac{\beta-r}{2r})^2\tau_p^2$, Recall that $a(t_p^{\pm})=(1\pm\eta_0)a(t_0)$. By the continuity of I and III on $a(t)$, if we choose η_0 sufficiently small, then for all $t_p^- \le t \le t_p^+$,

$$\widetilde{W}_0''(t)/\widetilde{W}_0(t) \le -\frac{3r-\beta}{4(\beta-r)}\cdot(\frac{\beta-r}{2r})^2\tau_p^2,$$

and the claim follows. □

5.14. *Proof of Lemma 3.1.* The following lemma is proved in Section 5.15.

LEMMA 5.10. *As $p\to\infty$, there is a constant $C>0$ such that with probability at least $1-o(1/p^3)$,*

$$\frac{\sqrt{p}|\widetilde{F}_p(t)-\widetilde{F}(t)|}{\sqrt{\widetilde{F}(t)(1-\widetilde{F}(t))}} \le \begin{cases} CK_p^3(\log(p))^{5/2}, & \forall\, 0<t<\sqrt{2\log(p)}, \frac{p}{2}>p\widetilde{F}(t)\ge\log^{5/4}(p), \\ CK_p^3\big(\log(p)\big)^{15/4}, & \forall\, 0<t<\sqrt{2\log(p)}, p\widetilde{F}(t)<\log^{5/4}(p). \end{cases}$$

We now prove Lemma 3.1. Put an evenly spaced grid on $[0,\sqrt{2\log p}]$ by $t_k=(\sqrt{2\log p}/p^2)k$, $0\le k\le p^2$. Denote by $V(t)=\sqrt{p}(\widetilde{F}_p(t)-\widetilde{F}(t))(\widetilde{F}(t)(1-\widetilde{F}(t)))^{-1/2}$. For each $0\le i\le p^2-1$, we claim that

$$\sup_{\{t_i\le t\le t_{i+1}\}}|V(t)| \le \max\{|V(t_i)|,|V(t_{i+1})|\}+L_p/p. \tag{5.56}$$

In fact, as both $\widetilde{F}_p(t)$ and $\widetilde{F}(t)$ are monotone functions, we have

$$\frac{\widetilde{F}_p(t_{i+1})-\widetilde{F}(t_i)}{\sqrt{\widetilde{F}(t_i)}} \le \frac{\widetilde{F}_p(t)-\widetilde{F}(t)}{\sqrt{\widetilde{F}(t)}} \le \frac{\widetilde{F}_p(t_i)-\widetilde{F}(t_{i+1})}{\sqrt{\widetilde{F}(t_{i+1})}}.$$

Let $h_i = \frac{\widetilde{F}(t_{i+1})}{\widetilde{F}(t_i)}$. Since $\widetilde{F}(t) \le \frac{1}{2}$, $\sup_{\{t_i \le t \le t_{i+1}\}}\{|V(t)|\}$ does not exceed

(5.57)
$$2\Big(\max\{\sqrt{\frac{1}{h_i}}|V(t_i)|, \sqrt{h_i}|V(t_{i+1})|\} + \frac{\sqrt{p}|\widetilde{F}(t_i)-\widetilde{F}(t_{i+1})|}{\sqrt{\widetilde{F}(t_i)}} + \frac{\sqrt{p}|\widetilde{F}(t_i)-\widetilde{F}(t_{i+1})|}{\sqrt{\widetilde{F}(t_{i+1})}}\Big).$$

Since the derivative of $(-\widetilde{F}(t))$ is the density of a location normal mixture, and is therefore bounded from above. Moreover, for $0 < t < \sqrt{2\log p}$ and sufficiently large p, $\widetilde{F}(t) \ge \widetilde{F}(\sqrt{2\log p}) \ge 2(1-K_p\epsilon_p)\bar{\Phi}(\sqrt{2\log p}) \ge p^{-1}L_p$. Using Taylor expansion,

(5.58)
$$\frac{\sqrt{p}|\widetilde{F}(t_i)-\widetilde{F}(t_{i+1})|}{\sqrt{\widetilde{F}(t_i)}} + \frac{\sqrt{p}|\widetilde{F}(t_i)-\widetilde{F}(t_{i+1})|}{\sqrt{\widetilde{F}(t_{i+1})}} \le \frac{L_p}{\sqrt{p^3\widetilde{F}(t_i)}} + \frac{L_p}{\sqrt{p^3\widetilde{F}(t_{i+1})}} \le L_p/p.$$

Similarly, we can show $|h_i - 1| \le L_p/p$. Inserting this and (5.58) into (5.57) gives (5.56).

Combining (5.56) with Lemma 5.10, the claim follows from

$$\sup_{\{0\le t\le\sqrt{2\log(p)}\}}\Big[\frac{\sqrt{p}|\widetilde{F}_p(t)-\widetilde{F}(t)|}{\sqrt{\widetilde{F}(t)(1-\widetilde{F}(t))}}\Big] = \sup_{\{0\le t\le\sqrt{2\log(p)}\}}|V(t)| \le C\sup_{\{0\le i\le p^2\}}|V(t_i)| + \frac{L_p}{p},$$

where $C > 0$ is some constant. □

5.15. *Proof of Lemma 5.10.* The following lemma is proved in Section 5.16.

LEMMA 5.11. *There are partitions* $\{1,2,\ldots,p\} = R_1' \cup R_2' \ldots \cup R_{N_1}' = R_1'' \cup R_2'' \ldots \cup R_{N_2}''$ *such that* $N_1 \le CK_p\log(p)$, $N_2 \le CK_p^2\log(p)$, *and that for any fixed* $1 \le j \le N_1$ *and* $1 \le k \le N_2$, *the collection of random variables* $\{\tilde{Z}(i)-\tilde{\mu}(i),\ i \in R_j'\}$ *are independent of each other, and the same are* $\{\tilde{\mu}(i),\ i \in R_k''\}$.

We now show Lemma 5.10. The key idea is to combine Lemma 1.1 with the well-known Bennett's inequality (e.g., [37]). The Bennett's inequality only applies to sum of independent random variables. To apply it in the

current setting, note that by Lemma 5.11, we can partition $\{1, 2, \ldots, p\}$ into N different subsets $R_1, \ldots, R_N$, where $N \leq CK_p^3 \log^2(p)$, such that the collection of random variables $\{\tilde{Z}(i) : i \in R_k\}$ are independent, for each $1 \leq k \leq N$. In light of this, we write $\widetilde{F}_p(t) = \frac{1}{p}\sum_{k=1}^N S_p^{(k)}(t)$, where $S_p^{(k)}(t) = \sum_{i \in R_k} 1\{|\tilde{Z}(i)| \geq t\}$ is the sum of independent random variables, to which the Bennet's inequality can be applied directly.

In detail, let $S^{(k)}(t) = E[S_p^{(k)}(t)]$ and $s_k = |R_k|$, $1 \leq k \leq N$, and $S(t) = \sum_{k=1}^N S^{(k)}(t)$. Since we are only interested in the region of t such that $\widetilde{F}(t) \leq 1/2$, it follows easily that

$$\frac{\sqrt{p}|\widetilde{F}_p(t) - \widetilde{F}(t)|}{\sqrt{\widetilde{F}(t)(1-\widetilde{F}(t))}} \lesssim \frac{\sqrt{2}|S_p(t) - S(t)|}{\sqrt{S(t)}} \leq \sum_{k=1}^N \frac{\sqrt{2}|S_p^{(k)}(t) - S^{(k)}(t)|}{\sqrt{S(t)}}. \tag{5.59}$$

For each $1 \leq k \leq N$, using Bennet's inequality [37, Page 851] yields

$$P\big(\big|S_p^{(k)} - S^{(k)}(t)\big| \geq \lambda\big) \leq 2\exp\Big(-\frac{\lambda^2}{2s_k\sigma_k^2}\psi\big(\frac{\lambda}{s_k\sigma_k^2}\big)\Big), \tag{5.60}$$

where ψ is as in [37, Page 851] and $s_k\sigma_k^2 = \mathrm{Var}(S_p^{(k)}(t))$. First, note that $x\psi(x)$ is monotonely increasing in $x \in (0, \infty)$. Second, by definitions and basic property of Bernoulli random variables, $s_k\sigma_k^2 \leq S^{(k)}(t) \leq S(t)$. Inserting these into (5.60) gives

$$P\Big(\big[S_p^{(\ell)} - S^{(\ell)}(t)\big] \geq \lambda\Big) \leq \exp\Big(-\frac{\lambda^2}{2S(t)}\psi\big(\frac{\lambda}{S(t)}\big)\Big).$$

Let $\lambda = C\sqrt{(\log p)S(t)}$ if $S(t) \geq \frac{1}{2}(\log p)^{5/4}$ and $\lambda = C(\log p)^{3/2}$ if $S(t) < \frac{1}{2}(\log p)^{5/4}$, where $C > 0$ is a constant. By elementary calculus and the property of ψ,

$$P\Big(\big[S_p^{(\ell)} - S^{(\ell)}(t)\big] \geq \lambda\Big) \leq \begin{cases} \exp\big(-\frac{C^2\log p}{2}\big), & S(t) \geq \frac{1}{2}(\log p)^{5/4} \\ \exp\big(-\frac{C\log p}{2}\big), & S(t) < \frac{1}{2}(\log p)^{5/4}. \end{cases}$$

Inserting this into (5.59) and noting that $p\widetilde{F}(t) \geq (\log p)^{-1/2}$ give the claim. □

5.16. *Proof of Lemma 5.11.* Recall that $\tilde{Z} - \tilde{\mu} \sim N(0, \Omega)$, the first claim follows directly from Lemma 1.1. For the second claim, introduce a graph $\mathcal{G} = (V, E)$ where $V = \{1, 2, \ldots, p\}$, and nodes i and j are connected if and only if $S_i \cap S_j = \emptyset$, where $S_i = \{1 \leq k \leq p : \Omega(i,k) \neq 0\}$, $1 \leq i \leq p$. Since Ω is K_p-sparse, $\mathcal{G}$ is K_p^2-sparse. Also, $\tilde{\mu}(i)$ and $\tilde{\mu}(j)$ are independent if and only if nodes i and j are disconnected. Applying Lemma 1.1 to $\mathcal{G}$ gives the claim. □

5.17. *Proof of Lemma 3.3.* Recall that $n_p = p^\theta$, $\hat{Z} = \hat{\Omega}Z$, and $\tilde{Z} = \Omega Z$. A direct result of Lemma 3.2 is that there is a term $0 < \eta_p \le CK_p^3(\log p)p^{-\theta/2}$ such that with probability at least $1 - o(1/p)$,

$$|1\{|\hat{Z}(j)| \ge t\} - 1\{|\tilde{Z}(j)| \ge t\}| \le 1\{t - \eta_p \le |\tilde{Z}(j)| < t + \eta_p\},\ \forall t > 0 \text{ and } 1 \le j \le p.$$

Let $G_p(t) = \widetilde{F}_p(t - \eta_p) - \widetilde{F}_p(t + \eta_p)$ and $G(t) = \widetilde{F}(t - \eta_p) - \widetilde{F}(t + \eta_p)$. By the above inequality, it is seen that with probability at least $1 - o(1/p)$,

$$|\bar{F}_p(t) - \widetilde{F}_p(t)| \le G_p(t). \tag{5.61}$$

We now analyze $G_p(t)$. By definitions and the triangle inequality,

$$G_p(t) \le G(t) + |\widetilde{F}_p(t - \eta_p) - \widetilde{F}(t - \eta_p)| + |\widetilde{F}_p(t + \eta_p) - \widetilde{F}(t + \eta_p)|. \tag{5.62}$$

A key fact is that there is a universal constant $C > 0$ such that

$$|\widetilde{F}'(t)| \le C(K_p\tau_p + t)\widetilde{F}(t). \tag{5.63}$$

To see the point, we write $\widetilde{F}(t) = \frac{1}{p}\sum_{i=1}^p E[\bar{\Psi}_{\sqrt{n_p}\tilde{\mu}(i)}(t)]$ and $\widetilde{F}'(t) = -\frac{1}{p}\sum_{i=1}^p E[\phi(t - \sqrt{n_p}\tilde{\mu}(i)) + \phi(t + \sqrt{n_p}\tilde{\mu}(i))]$, where ϕ is the density function of $N(0,1)$. Note that there is a constant $C > 0$ such that $\phi(x) \le C|x|\bar{\Phi}(x)$, and that $|t \pm \sqrt{n_p}\tilde{\mu}(i)| \le t + K_p\tau_p$ for all $1 \le i \le p$, the desired claim follows.

Now, first, write $G(t) = \widetilde{F}(t-\eta_p) - \widetilde{F}(t+\eta_p) = 2\eta_p\widetilde{F}'(\xi)$ for some number ξ with $|\xi - t| < \eta_p$. Using (5.63), $|\widetilde{F}'(\xi)| \le CK_p\tau_p\widetilde{F}(\xi) \sim CK_p\tau_p\widetilde{F}(t)$. It follows

$$G(t) \le CK_p\tau_p\widetilde{F}(t)\eta_p. \tag{5.64}$$

Second, by Lemma 3.1 and monotonicity, with probability at least $1 - o(1/p)$, $|\widetilde{F}_p(t \pm \eta_p) - \widetilde{F}(t \pm \eta_p)| \le CK_p^3(\log p)^{7/2}p^{-1/2}(\widetilde{F}(t \pm \eta_p))^{1/2}$, where by (5.63), $\widetilde{F}(t \pm \eta_p) \asymp \widetilde{F}(t)$. It follows that with probability at least $1 - o(1/p)$,

$$|\widetilde{F}_p(t \pm \eta_p) - \widetilde{F}(t \pm \eta_p)| \le CK_p^3(\log p)^4p^{-1/2}(\widetilde{F}(t))^{1/2}. \tag{5.65}$$

Recall that $\eta_p \le K_p^3(\log p)p^{-\theta/2}$. Inserting (5.64)-(5.65) into (5.62) gives

$$G_p(t) \le CK_p^4(\log p)^{3/2}p^{-\theta/2}\widetilde{F}(t) + CK_p^3(\log p)^4p^{-1/2}(\widetilde{F}(t))^{1/2}. \tag{5.66}$$

Combining (5.66) with (5.61) gives

$$\frac{\sqrt{p}|\bar{F}_p(t) - \widetilde{F}_p(t)|}{\sqrt{\widetilde{F}(t)(1 - \widetilde{F}(t))}} \le \frac{\sqrt{p}|G_p(t)|}{\sqrt{\widetilde{F}(t)(1 - \widetilde{F}(t))}} \le C\big(K_p^4(\log p)^{3/2}(p^{1-\theta}\widetilde{F}(t))^{1/2} + K_p^3(\log p)^4\big), \tag{5.67}$$

and the claim follows. □

5.18. *Proof of Theorem 3.1.* We consider the case when $p\widetilde{F}(t) < K_p^6(\log(p))^9$ and when $p\widetilde{F}(t) \geq K_p^6(\log(p))^9$ separately.

In the first case, it is sufficient to show that $|HC(t, \bar{F}_p)| \leq L_p$ and $|HC(t, \widetilde{F}_p)| \leq L_p$. By Lemmas 3.3 and 5.10, with probability at least $1 - o(1/p)$, $p|\bar{F}_p(t) - \widetilde{F}(t)| \leq L_p$. By Lemma 5.4, $\widetilde{F}(t) \geq (1 - K_p\epsilon_p)\bar{\Psi}(t)$ and thus, $p\bar{\Psi}(t) \leq L_p$. Since $HC(t, \bar{F}_p)$ is defined in a way such that $\bar{F}_p(t) \geq 1/p$, it is easy to see that $HC(t, \bar{F}_p) \leq p|\bar{F}_p(t) - \bar{\Psi}(t)| \leq p|\bar{F}_p(t) - \widetilde{F}(t)| + p\widetilde{F}(t) + p\bar{\Psi}(t) \leq L_p$. Similarly, we can prove that $HC(t, \widetilde{F}_p) \leq L_p$. The claim follows easily.

In the second case, let $h(t) = (\widetilde{F}(t)(1-\widetilde{F}(t)))/(\bar{F}_p(t)(1-\bar{F}_p(t))$ and write for short $g(t) = \sqrt{p}\big(\bar{F}_p(t) - \widetilde{F}_p(t)\big)\big(\widetilde{F}(t)(1 - \widetilde{F}(t)\big)^{-1/2}$. By definitions, we can write

$$HC(t, \bar{F}_p) - HC(t, \widetilde{F}) = g(t)\sqrt{h(t)} + HC(t, \widetilde{F})(\sqrt{h(t)} - 1). \tag{5.68}$$

We first prove $|h(t) - 1| \leq o(1)$. To see this, note that (5.67) and Lemma 5.10 ensure that with probability at least $1 - o(1/p)$,

$$\begin{aligned} |\bar{F}_p(t)/\widetilde{F}(t) - 1| &\leq |(\bar{F}_p(t) - \widetilde{F}_p(t))/\widetilde{F}(t)| + |\widetilde{F}_p(t)/\widetilde{F}(t) - 1| \\ &\leq CK_p^4(\log p)^{3/2}p^{-\theta/2} + CK_p^3(\log p)^4(p\widetilde{F}(t))^{-1/2}. \end{aligned} \tag{5.69}$$

By the assumption of $p\widetilde{F}(t) \geq K_p^6(\log p)^9$, the right hand side of (5.69) tends to 0. Thus, with probability at least $1-o(1/p)$, $0 \leq \bar{F}_p(t), \widetilde{F}(t) < 2/3$ for all $0 < t \leq \sqrt{2\log(p)}$. Note that for all $x, y \in (0, 2/3)$, $|[x(1-x)]/[y(1-y)]-1| \leq C|x/y - 1|$. It follows from (5.69) and definitions that

$$|h(t) - 1| \leq C|\bar{F}_p(t)/\widetilde{F}(t) - 1| \leq L_p(p^{-\theta/2} + (p\widetilde{F}(t))^{-1/2}), \tag{5.70}$$

where the right hand side tends to 0 since $p\widetilde{F}(t) \geq K_p^6(\log p)^9$. At the same time, since $\widetilde{F}(t) \geq (1 - K_p\epsilon_p)\bar{\Psi}(t)$, we have $|\widetilde{F}(t) - \bar{\Psi}(t)| \leq \widetilde{F}(t) + \bar{\Psi}(t) \lesssim 2\widetilde{F}(t)$. It follows from $1 - \widetilde{F}(t) \geq 1 - \bar{\Psi}(t) - K_p\epsilon_p \geq 1/2 - K_p\epsilon_p$ that

$$|HC(t, \widetilde{F})| = \sqrt{p}|\widetilde{F}(t) - \bar{\Psi}(t)|(\widetilde{F}(t)(1 - \widetilde{F}(t))^{-1/2} \leq C(p\widetilde{F}(t))^{1/2}. \tag{5.71}$$

Combining (5.70) and (5.71) gives

$$HC(t, \widetilde{F})|\sqrt{h(t)} - 1| \leq L_p[(p^{1-\theta}\widetilde{F}(t))^{1/2} + 1]. \tag{5.72}$$

At the same time, a direct use of Lemma 3.3 also gives that with probability at least $1 - o(1/p)$,

$$g(t) \leq L_p[(p^{1-\theta}\widetilde{F}(t))^{1/2} + 1]. \tag{5.73}$$

Inserting (5.72) and (5.73) into (5.68) and recalling $|h(t) - 1| \to 0$ gives the claim. □

5.19. *Proof of Theorem 3.2.* Write for short $\widehat{W}_p(t) = p^{-1/2}HC(t, \bar{F}_p)$ and $W(t) = p^{-1/2}HC(t, \widetilde{F}_p)$.

First consider the case of $\theta \geq \frac{1}{2}$. By triangle inequality, Theorem 3.1, and Lemma 2.3 we have

(5.74)
$$\sup_{\bar{\Psi}^{-1}(\frac{1}{2})<t<s_p^*} |\widehat{W}_p(t) - W_0(t)| \leq \sup_{\bar{\Psi}^{-1}(\frac{1}{2})<t<s_p^*} |\widehat{W}_p(t) - W(t)| + \sup_{t>\bar{\Psi}^{-1}(\frac{1}{2})} |W(t) - W_0(t)|$$
$$\leq L_p(p^{-\beta} + p^{-\theta/2}\sqrt{\widetilde{F}(t)} + p^{-1/2}) \leq L_p(p^{-\theta/2} + p^{-\beta}).$$

This result is parallel to Lemma 2.3. When $r < \beta$, similar to (5.51) we can obtain that for all u satisfying $|u| \leq c_4/\tau_p$,

(5.75)
$$\widehat{W}_p(t_p^{**} + u) - \widehat{W}_p(t_p^{**}) \leq L_p(p^{-\frac{\theta}{2}} + p^{-\beta}) + [L_p p^{-c_0(\beta,r,a)} - \frac{c_7 u^2}{1+u^2}] \sup_{\{t\geq 0\}} \widetilde{W}_0(t),$$

for some constant $c_7 > 0$, where t_p^{**} is as in (5.51). It is easy to check that $\sup_{\{t\geq 0\}} \widetilde{W}_0(t) = L_p p^{-\delta(\beta,r)} > p^{-\frac{1-\theta}{2}} > p^{-\theta/2}$, $c_0(\beta, r, a, \theta) < \beta$, and $p^{-c_0(\beta,r,a)} \sup_{t\geq 0} \widetilde{W}_0(t) \geq p^{-\beta}$. Thus, for any $u > L_p p^{-c_2(\beta,r,a)}$ with $c_2(\beta, r, a) < \min\{\frac{\theta-2\delta(\beta,r)}{4}, \frac{c_0(\beta,r,a)}{2}\}$, it holds that $\widehat{W}_p(t_p^{**}+u) - \widehat{W}_p(t_p^{**}) = -L_p p^{-2c_2(\beta,r,a)}(1 + o(1)) < 0$ for all $|u| \leq c_4/\tau_p$. Again, using similar arguments as in Theorem 2.2, we can prove that $\widehat{W}_p(t) - \widehat{W}_p(t_p^{**}) < 0$ for all $|t - t_p^{**}| > c_4/\tau_p$. Thus, we have proved that

$$|t_p^{HC} - t_p^{**}| = |T_{HC}(\bar{F}_p) - t_p^{**}| \leq L_p p^{-c_2(\beta,r,a)}.$$

This together with (5.52) completes the proof of the Theorem when $r < \beta$.

Now we consider the case where $r \geq \beta$. If $t > \tau_p$ or $t < \sqrt{2\beta \log p} - \Delta_1$ with $\Delta_1 = d_0 \log\log p/\sqrt{\log p}$, by Lemma 2.4 and (5.74), it holds $\widehat{W}_p(t) = W_0(t) + (\widehat{W}_p(t) - W_0(t)) \lesssim \frac{1}{\sqrt{2}} p^{-\beta/2} + L_p p^{-\beta} + L_p p^{-\theta/2}$. Recall that $\beta < 1 - \theta \leq \theta$. Thus $\widehat{W}_p(t) \lesssim \frac{1}{\sqrt{2}} p^{-\beta/2}(1 + o(1))$. If $\sqrt{2\beta \log p} - \Delta_1 < t < \tau_p$, using similar argument we obtain that $\widehat{W}_p(t) = W_0(t) + (\widehat{W}_p(t) - W_0(t)) \gtrsim p^{-\beta/2}(1 - o(1))$. Thus,

$$t_p^{HC} \in (\sqrt{2\beta \log p} - \Delta_1, \tau_p)$$

and the claim in the theorem follows.

Next we consider the case where $\theta < \frac{1}{2}$. By Theorem 3.1 and Lemma 2.4 and noting that $1-\theta > \beta > \frac{1-\theta}{2}$, for any $t, t+u \in [s_p(\theta), s_p^*]$ we have

$$\widehat{W}_p(t+u) - \widehat{W}_p(t) = (\widehat{W}_p(t+u) - W_0(t+u)) - (\widehat{W}_p(t) - W_0(t)) + (W_0(t+u) - W_0(t)) \le L_p p^{-\theta/2}\sqrt{\widetilde{F}(t)} + L_p p^{-\beta} + (W_0(t+u) - W_0(t)).$$

Since $p^{-\theta}\widetilde{F}(t) \le p^{-1+\theta}$ and $\beta > (1-\theta)/2$, it follows that

$$\widehat{W}_p(t+u) - \widehat{W}_p(t) \le L_p p^{-(1-\theta)/2} + L_p p^{-\beta} + (W_0(t+u) - W_0(t)). \tag{5.76}$$

So the stochastic behavior of $W_0(t)$ in the range $t \in [s_p(\theta), s_p^*]$ determines the stochastic behavior of $\widehat{W}_p(t+u) - \widehat{W}_p(t)$. By direct calculations, we obtain that if (β, r, θ) falls in either of the six sub-regions as follows

- $1/3 < \theta \le 1/2$, $(1-\theta)/2 < \beta < 1-\theta$, $r > \max\{\rho_\theta^*(\beta), \frac{1-2\theta}{4}\}$,
- $\frac{1}{4} < \theta \le \frac{1}{3}$, $(1-\theta)/2 < \beta \le 1-2\theta$, $r > \max\{\frac{1-2\theta}{4}, \rho_\theta^*(\beta)\}$, $|r - \sqrt{1-2\theta}| \ge \sqrt{1-2\theta-\beta}$
- $\frac{1}{4} < \theta \le \frac{1}{3}$, $1-2\theta < \beta \le 1-\theta$, $r > \max\{\frac{1-2\theta}{4}, \rho_\theta^*(\beta)\}$
- $0 < \theta \le \frac{1}{4}$, $(1-\theta)/2 < \beta \le 3(1-2\theta)/4$, $r > \max\{\frac{\beta}{3}, \rho_\theta^*(\beta)\}$, $|r - \sqrt{1-2\theta}| \ge \sqrt{1-2\theta-\beta}$
- $0 < \theta \le \frac{1}{4}$, $3(1-2\theta)/4 < \beta \le 1-2\theta$, $r > \max\{\frac{1-2\theta}{4}, \rho_\theta^*(\beta)\}$, $|r - \sqrt{1-2\theta}| \ge \sqrt{1-2\theta-\beta}$
- $0 < \theta \le \frac{1}{4}$, $1-2\theta < \beta < 1-\theta$, $r > \max\{\frac{1-2\theta}{4}, \rho_\theta^*(\beta)\}$,

then $t_p^{**} \in (s_p(\theta), s_p^*)$ and the maximum of $W_0(t)$ is achieved in $(s_p(\theta), s_p^*)$. So it reduces to the $\theta > 1/2$ case. Note that the six regions above can be summarized into Condition (a)-(b) in Theorem 1.3. By (5.76) and using similar proof as that for $\theta \ge \frac{1}{2}$ we finish the proof of Theorem 3.2. □

5.20. *Proof of Lemma 3.4.* Introduce $u_p(t) = u_p(t, \epsilon_p, \tau_p, \Omega) = \sum_{j=1}^p E\big[\tilde{\mu}(j)^2 \cdot 1\{|\tilde{Z}(j)| \ge t\}\big]$. The following lemma is proved in the appendix.

LEMMA 5.12. *For any $t > 0$, there are universal constants $C_1 > 0$ and $C_2 > 0$ such that for sufficiently large p, $C_1 \min\{t, \frac{1}{K_p\sqrt{2\log p}}\}\sqrt{n_p} \le \frac{m_p(t,\epsilon_p,\tau_p,\Omega)}{u_p(t,\epsilon_p,\tau_p,\Omega)} \le C_2(1+t)\sqrt{n_p}$ and $m_p(t, \epsilon_p, \tau_p, \Omega) \le C_2(1+t)K_p^2\tau_p^2 n_p^{-1/2} p\widetilde{F}(t)$, where $\widetilde{F}(t)$ is defined in Lemma 5.2.*

The following lemma is proved in Section 5.21.

LEMMA 5.13. *There is a constant $C > 0$ such that with probability at least $1 - o(1/p)$, for all $0 \le t \le \sqrt{2\log(p)}$,*

$$\sqrt{n_p}|M_p(t, \tilde{Z}, \mu) - m_p(t, \epsilon_p, \tau_p, \Omega)| \le CK_p^5(\log p)^4\sqrt{p\widetilde{F}(t)}, \tag{5.77}$$

$$|V_p(t, \tilde{Z}, \mu) - v_p(t, \epsilon_p, \tau_p, \Omega)| \le CK_p^4(\log p)^{9/2}\sqrt{p\widetilde{F}(t)}. \tag{5.78}$$

Write for short $\tilde{V}_p(t) = V_p(t, \tilde{Z}, \mu)$, $\tilde{M}_p(t) = M_p(t, \tilde{Z}, \mu)$, $m_p(t) = m_p(t, \epsilon_p, \tau_p, \Omega)$, $v_p(t) = v_p(t, \epsilon_p, \tau_p, \Omega)$, $\widetilde{Sep}(t) = \widetilde{Sep}(t, \epsilon_p, \tau_p, \Omega)$, $Sep(t) = Sep(t, \tilde{Z}, \mu, \Omega)$, $\widetilde{F}(t) = \widetilde{F}(t, \epsilon_p, \tau_p, \Omega)$ and $\widetilde{F}_p(t) = \widetilde{F}_p(t, \tilde{Z}, \mu, \Omega)$. We consider the two cases 1) $t > \tau_p + \tilde{s}_p$ or $p\widetilde{F}(t) \le K_p^8(\log p)^{10}$, and 2) $t \le \tau_p + \tilde{s}_p$ and $p\widetilde{F}(t) > K_p^8(\log p)^{10}$, separately, where $\tilde{s}_p$ is defined in Lemma 5.4.

Consider the first case. It suffices to show (1a) $p^{(\theta-1)/2}\widetilde{Sep}(t) \le L_p p^{-1/2} + L_p p^{-\max\{4\beta-2r, 3\beta+r\}/4}$ and (1b) $p^{(\theta-1)/2}Sep(t) \le L_p p^{-\max\{4\beta-2r,3\beta+r\}/4} + L_p p^{-1/2}$. Claim (1a) can be proved using the same arguments as in Lemma 2.1, so we only need to prove (1b).

Consider (1b). Let η be a $p \times 1$ vector such that $\eta(j) = 1\{(\Omega\hat{\mu}_t^{\tilde{Z}})(j) \ne 0\}$, $1 \le j \le p$. Also, for any $p \times 1$ vectors x and y, let $x \circ y$ be the $p \times 1$ vector such that $(x \circ y)(j) = x(j)y(j)$, $1 \le j \le p$. By definitions, it is seen that $\tilde{M}_p(t) = (\hat{\mu}_t^{\tilde{Z}})'\Omega\mu = (\hat{\mu}_t^{\tilde{Z}})'\Omega(\mu \circ \eta)$. Using Cauchy-Schwartz inequality, $|\tilde{M}_p(t)| \le \big((\hat{\mu}_t^{\tilde{Z}})'\Omega\hat{\mu}_t^{\tilde{Z}}\big)^{1/2}\big((\mu \circ \eta)'\Omega(\mu \circ \eta)\big)^{1/2}$. Recalling that $\tilde{V}_p(t) = (\hat{\mu}_t^{\tilde{Z}})'\Omega\hat{\mu}_t^{\tilde{Z}}$, it follows that

$$|Sep(t)| = 2|\tilde{M}_p(t)|(V_p(t))^{-1/2} \le 2\big((\mu \circ \eta)'\Omega(\mu \circ \eta)\big)^{1/2}.$$

Since the largest eigenvalue of Ω is no greater than K_p, the last term above $\le 2K_p^{1/2}\|\mu \circ \eta\|$ and so $|Sep(t)| \le 2K_p^{1/2}\|\mu \circ \eta\|$. At the same time, by Lemma 3.1, with probability at least $1 - o(1/p)$, $p\widetilde{F}_p(t) \le p|\widetilde{F}_p(t) - \widetilde{F}(t)| + p\widetilde{F}(t) \le L_p(p\widetilde{F}(t))^{1/2} + p\widetilde{F}(t) \le L_p p^{1-\max\{4\beta-2r,3\beta+r\}/2}$ if $t \ge \tau_p + \tilde{s}_p$. Similarly, we can show that $p\widetilde{F}_p(t) \le L_p$ if $p\widetilde{F}(t) \le K_p^8(\log p)^{10}$. Thus, in case (1b) we have $p\widetilde{F}(t) \le L_p p^{1-\max\{4\beta-2r,3\beta+r\}/2} + L_p$. By definitions, this implies that $\hat{\mu}_t^{\tilde{Z}}$ has no more than $L_p p^{1-\max\{4\beta-2r,3\beta+r\}/2} + L_p$ non-zero coordinates. Since Ω is K_p-sparse, η also has no more than $L_p p^{1-\max\{4\beta-2r,3\beta+r\}/2} + L_p$ nonzero coordinates. Therefore, $\|\mu \circ \eta\| \le L_p p^{\frac{1-\theta}{2}-\max\{4\beta-2r,3\beta+r\}/4} + L_p p^{-\theta/2}$, and (1b) follows from the assumption that $K_p \le L_p$.

Consider the second case. Denote $h(t) = v_p(t)/\tilde{V}_p(t)$. The key is to show

$$|h(t) - 1| \le L_p(p\widetilde{F}(t))^{-1/2}. \tag{5.79}$$

Towards this end, we write $|h(t)-1| = I \cdot II \cdot h(t) \cdot (p\widetilde{F}(t))^{-1/2}$, where $I = |\tilde{V}_p(t) - v_p(t)|(p\widetilde{F}(t))^{-1/2}$, and $II = (p\widetilde{F}(t))/v_p(t)$. First, by Lemma 5.13, $I \le L_p$ with probability at least $1 - o(1/p)$. Second, by Lemma 5.4, $II \le C$ with some constant $C > 0$ whose value depends on whether $r < \beta$ and $t \le \tau_p + \tilde{s}_p$ or $r \ge \beta$. Last, by Lemma 5.13 and (5.78), with probability at least $1 - o(1/p)$, $\tilde{V}_p(t)/v_p(t) \ge 1 - CK^4(\log p)^{9/2}\frac{(p\widetilde{F}(t))^{1/2}}{v_p(t)} \ge 1 - o(1)$, where we note that $p\widetilde{F}(t) \ge K_p^8(\log p)^9$ and $CK_p^4(\log(p))^{9/2}(p\widetilde{F}(t))^{1/2}(v_p(t))^{-1} \lesssim K_p^4(\log(p))^{9/2}(p\widetilde{F}(t))^{-1/2} = o(1)$. As a result, with probability at least $1 - o(1/p)$, $h(t) = \frac{\tilde{V}_p(t)}{v_p(t)} \lesssim 1$. Combining these gives (5.79).

Next, write

$$|Sep(t) - \widetilde{Sep}(t)| = |\frac{\tilde{M}_p(t)}{\sqrt{\tilde{V}_p(t)}} - \frac{m_p(t)}{\sqrt{v_p(t)}}| \le III + IV, \tag{5.80}$$

where $III = |\tilde{M}_p(t) - m_p(t)|\sqrt{h(t)}/\sqrt{v_p(t)}$ and $IV = m_p(t)|\sqrt{h(t)} - 1|/\sqrt{v_p(t)}$. Recall that $h(t) \lesssim 1 + L_p$ and that $Cp\widetilde{F}(t) \le v_p(t)$. It follows from Lemma 5.13 that with probability at least $1 - o(1/p)$, $III \lesssim |\tilde{M}_p(t) - m_p(t)|(p\bar{F}_p(t))^{-1/2} \le L_p n_p^{-1/2}$. At the same time, note that $IV \le |h(t) - 1|m_p(t)(v_p(t))^{-1/2}$. On one hand, by Lemmas 5.4 and 5.12, $m_p(t) \le L_p n_p^{1/2} u_p(t) \le L_p K_p^2 n_p^{-1/2} p\widetilde{F}(t)$. On the other hand, since $v_p(t) \ge Cp\widetilde{F}(t)$, by (5.79), we have $IV \le L_p n_p^{-1/2}$ with probability at leats $1 - o(1/p)$. Combining these with (5.80) gives the claim.

By going through the proof above we see that if further $\Omega \in \widetilde{M}_p^*(a, b, K_p)$, then the two cases at the very beginning can be reduced to 1) $p\widetilde{F}(t) \le K_p^8(\log p)^{10}$, and 2) $p\widetilde{F}(t) > K_p^8(\log p)^{10}$, and the claim $|Sep(t) - \widetilde{Sep}(t)| \le L_p n_p^{-1/2}$ can be proved using same arguments. Thus, Lemma 3.4 is proved. □

5.21. *Proof of Lemma 5.13.* Write for short $\tilde{M}_p(t) = M_p(t, \tilde{Z}, \mu)$, $\tilde{V}_p(t) = V_p(t, \tilde{Z}, \mu)$, $m_p(t) = m_p(t, \epsilon_p, \tau_p, \Omega)$, and $v_p(t) = E[V_p(t, \tilde{Z}, \mu)]$. The following Lemma is proved in Section 5.22.

LEMMA 5.14. *For any* $t \in (0, \sqrt{2\log p}]$,

$$P\Big(\sqrt{n_p}|\tilde{M}_p(t) - m_p(t)| \ge CK_p^3(\log p)^2\lambda\Big) \le CK_p^3(\log p)^2 \exp\Big(-\frac{\lambda^2 c_2}{2K_p\sqrt{2\log(p)n_p}m_p(t)}\psi\big(\frac{\lambda c_2}{\sqrt{n_p}m_p(t)}\big)\Big),$$

$$P\Big(|\tilde{V}_p(t) - v_p(t)| \ge CK_p^3(\log p)^2\lambda\Big) \le CK_p^3(\log p)^2 \exp\left(-\frac{\lambda^2}{4K_p p\widetilde{F}(t)}\psi\Big(\frac{\lambda}{2K_p p\widetilde{F}(t)}\Big)\right),$$

where ψ *is as in Bennett's lemma [37, Page 851].*

Since the proofs are very similar, we only show the first one. The goal is to show that with probability $1-o(1/p^3)$, $|\tilde{M}_p(t)-m_p(t)| \le CK_p^4(\log(p))^{9/2}(p\widetilde{F}(t))^{-1/2}$ for any $0 \le t \le \sqrt{2\log(p)}$. Once this is shown, we lay out an evenly spaced grid on $[0, \sqrt{2\log(p)}]$ with an inter-distance of $1/p$, and the claim follows by similar argument as in the proof of Lemma 3.1.

Since Lemma 5.12 ensures that $m_p(t) \le CK_p^2(\log p)^{3/2}pn_p^{-1/2}\widetilde{F}(t)$, by the monotonicity of $x\psi(x)$ and Lemma 5.14,

(5.81)
$$P\Big(\sqrt{n_p}|\tilde{M}_p(t) - m_p(t)| \ge CK_p^3(\log p)^2\lambda\Big) \le CK_p^3(\log p)^2 \exp\Big(-\frac{\lambda^2 c_2}{2CK_p^3(\log p)^2 p\widetilde{F}(t)}\psi\big(\frac{\lambda c_2}{CK_p^2(\log p)^{3/2}p\widetilde{F}(t)}\big)\Big).$$

We now show the desired claim for the case $p\widetilde{F}(t) \ge (\log(p))^{3/2}$ and the case $p\widetilde{F}(t) \le (\log(p))^{3/2}$ separately.

Consider the first case. Let $\lambda = CK_p^2(\log p)^{3/2}\sqrt{p\widetilde{F}(t)}$. Direct calculations show that $\lambda/[K_p^2(\log p)^{3/2}p\widetilde{F}(t)] \le C(p\widetilde{F}(t))^{-1/2}$ and $\lambda^2/[K_p^3(\log p)^2 p\widetilde{F}(t)] \ge C\log(p)K_p$. By (5.81) and noting that $\lim_{x\to 0+}\psi(x) = 1$,

$$P\left(\sqrt{n_p}|\tilde{M}_p(t) - m_p(t)| \ge CK_p^5(\log p)^{7/2}\sqrt{p\widetilde{F}(t)}\}\right) \le CK_p^3(\log p)^2 \exp\Big(-\frac{C^2K_p(\log p)}{2}\Big) \le o(1/p^3).$$

Consider the second case. Let $\lambda = CK_p^2(\log p)^3$. It is seen that $\lambda/[K_p^2(\log p)^{3/2}p\widetilde{F}(t)] \ge C(\log(p))^{3/2}/(p\widetilde{F}(t))$. Using Lemma 5.14 where we note that $\psi(x) \sim \frac{\log(x)}{x}$ when $x \to \infty$ [37, Page 852],

$$P\Big(\sqrt{n_p}|\tilde{M}_p(t) - m_p(t)| \ge CK_p^5(\log p)^{7/2}\Big) \le C(\log p)^5 \exp\Big(-\frac{C(\log p)}{2}\Big) \le o(1/p^3).$$

This together with $p\widetilde{F}(t) \gtrsim p(1-K_p\epsilon_p)\bar{\Psi}(t) \gtrsim (\log p)^{-1/2}$ yields the desired claim. □

5.22. *Proof of Lemma 5.14.* Since the proofs are similar, we only show the first one. By Lemma 1.1, we can partition $\{1,\cdots,p\}$ into $N = N_1N_2 \le K_p^3\log^2(p)$ sets $R_1,\cdots,R_N$ such that for any fixed index $1 \le k \le N$, the collection of bivariate random variables $\{(\tilde{\mu}(j),\tilde{Z}(j)) : j \in R_k\}$ are independent of each other. Recall that $\tilde{M}_p(t) = \sum_{j=1}^p \tilde{\mu}(j)\mathrm{sgn}(\tilde{Z}(j))1\{|\tilde{Z}(j)| \ge t\}$ and $m_p(t) = E[\tilde{M}_p(t)]$. The partition allows us to write $\tilde{M}_p(t) - m_p(t) = \sum_{k=1}^N[\tilde{M}_p^{(k)}(t) - m_p^{(k)}(t)]$, where $M_p^{(k)}(t) = \sum_{j\in R_k}\tilde{\mu}(j)\mathrm{sgn}(\tilde{Z}(j))1\{|\tilde{Z}(j)| \ge t\}$ and $m_p^{(k)}(t) = E[M_p^{(k)}(t)]$, $1 \le k \le N$. It follows that for any $\lambda > 0$,

$$P(\sqrt{n_p}|\tilde{M}_p(t)-m_p(t)| \ge N\lambda) \le \sum_{k=1}^N P(\sqrt{n_p}|\tilde{M}_p^{(k)}(t)-m_p^{(k)}(t)| \ge \lambda). \tag{5.82}$$

Fix $1 \le k \le N$, using Bennett's inequality [37, Page 851],

$$P(\sqrt{n_p}|\tilde{M}_p^{(k)}(t) - m_p^{(k)}(t)| \ge \lambda) \le \exp\Big(-\frac{\lambda^2}{2|R_k|\sigma_k^2}\psi\Big(\frac{\lambda K_p\sqrt{2\log p}}{|R_k|\sigma_k^2}\Big)\Big),$$

where ψ is as in [37, Page 851], and $|R_k|\sigma_k^2$ is the variance of $\sqrt{n_p}\tilde{M}_p^{(k)}(t)$. Using Lemma 5.12, $|R_k|\sigma_k^2 \le n_pu_p(t) \le c_2^{-1}K_p\sqrt{2\log(p)n_p}m_p(t)$. By the monotonicity of the function $x\psi(x)$ [37, Page 851], it follows that

$$P\Big(\sqrt{n_p}|\tilde{M}_p^{(k)}(t)-m_p^{(k)}(t)| \ge \lambda\Big) \le \exp\Big(-\frac{\lambda^2c_2}{2K_p\sqrt{2\log(p)n_p}m_p(t)}\psi\big(\frac{\lambda c_2}{\sqrt{n_p}m_p(t)}\big)\Big).$$

Inserting this into (5.82), the claim follows by recalling $N \le CK_p^3\log^2(p)$. □

5.23. *Proof of Lemma 3.5.* Write for short $\hat{M}_p(t) = M_p(t,\hat{Z},\mu)$, $\tilde{M}_p(t) = M_p(t,\tilde{Z},\mu)$, $\hat{V}_p(t) = V_p(t,\hat{Z},\mu)$, and $\tilde{V}_p(t) = V_p(t,\tilde{Z},\mu)$, $m_p(t) = m_p(t,\epsilon_p,\tau_p,\Omega)$, and $v_p(t) = v_p(t,\epsilon_p,\tau_p,\Omega)$. We discuss the case 1) $t > \tau_p + \tilde{s}_p$ or $p\widetilde{F}(t) \le K_p^{10}(\log p)^{10}$ and the case 2) $t \le \tau_p+\tilde{s}_p$ and $p\widetilde{F}(t) > K_p^{10}(\log p)^{10}$ separately.

Consider the first case. First, in the proof of Lemma 3.4, we have shown that $Sep(t,\tilde{Z},\mu,\Omega) \le L_pp^{\frac{1-\theta}{2}-\frac{1}{4}\max\{4\beta-2r,3\beta+r\}} + L_pp^{-\theta/2}$. Second, by similar argument as in the proof Lemma 3.4 part (1b), and using Lemma 3.3, we can prove that $Sep(t,\hat{Z},\mu,\hat{\Omega}) \le L_pp^{\frac{1-\theta}{2}-\frac{1}{4}\max\{4\beta-2r,3\beta+r\}} + L_pp^{-\theta/2}$. Combining these gives the claim.

Consider the second case. The key is that with probability at least $1-o(1/p)$,

$$\max\{\sqrt{n_p}|\hat{M}_p(t)-\tilde{M}_p(t)|,\ |\hat{V}_p(t)-\tilde{V}_p(t)|\} \le L_p\cdot[p^{1-\theta/2}\widetilde{F}(t)+(p\widetilde{F}(t))^{1/2}], \tag{5.83}$$

$$\max\{|(\tilde{V}_p(t)-v_p(t))/v_p(t)|,\ |\sqrt{n_p}(\tilde{M}_p(t)-m_p(t))/v_p(t)|\}=o(1). \tag{5.84}$$

for all $t \le \tau_p+\tilde{s}_p$. To see (5.83), note that $|\hat{M}_p(t)-\tilde{M}_p(t)| = |(\hat{\mu}_t^{\hat{Z}}-\hat{\mu}_t^{\tilde{Z}})'\Omega\mu| \le \|\hat{\mu}_t^{\hat{Z}}-\hat{\mu}_t^{\tilde{Z}}\|_1\cdot\|\Omega\mu\|_\infty$, where by the K_p-sparsity of Ω, $\|\Omega\mu\|_\infty \le K_p\tau_p n_p^{-1/2}$, and so $|\hat{M}_p(t)-\tilde{M}_p(t)| \le K_p\tau_p n_p^{-1/2}\|\hat{\mu}_t^{\hat{Z}}-\hat{\mu}_t^{\tilde{Z}}\|_1$. Similarly, $\|\Omega(\hat{\mu}_t^{\hat{Z}}+\hat{\mu}_t^{\tilde{Z}})\|_\infty \le \|\Omega\|_1\|\hat{\mu}_t^{\hat{Z}}+\hat{\mu}_t^{\tilde{Z}}\|_\infty \le 2K_p$, and so $|\hat{V}_p(t)-\tilde{V}_p(t)| \le |(\hat{\mu}_t^{\hat{Z}}-\hat{\mu}_t^{\tilde{Z}})'\Omega(\hat{\mu}_t^{\hat{Z}}+\hat{\mu}_t^{\tilde{Z}})| \le 2K_p\|\hat{\mu}_t^{\hat{Z}}-\hat{\mu}_t^{\tilde{Z}}\|_1$. By similar argument as in the proof of Lemma 3.3, it is seen that with probability at least $1-o(1/p)$, $\|\hat{\mu}_t^{\hat{Z}}-\hat{\mu}_t^{\tilde{Z}}\|_1 \le pG_p(t)$, where $G_p(t)$ is defined therein. It is shown in Lemma 3.3 that $G_p(t) \le CK_p^4(\log p)^{3/2}p^{-\theta/2}\widetilde{F}(t)+CK_p^3(\log p)^4p^{-1/2}(\widetilde{F}(t))^{1/2}$ with probability at least $1-o(1/p)$. Combining these gives (5.83).

To see (5.84), note that by Lemma 5.13, with probability at least $1-o(1/p)$,

$$|\tilde{V}_p(t)-v_p(t)| \le CK_p^4(\log(p)^{9/2}(p\widetilde{F}(t))^{1/2}. \tag{5.85}$$

Recall that by Lemma 5.4, $v_p(t) \ge Cp\widetilde{F}(t)$ with some constant $C>0$ whose value depends on whether $r<\beta$ or $r\ge\beta$. Combining this with the fact that $p\widetilde{F}(t) \ge K_p^{10}(\log p)^{10}$ for all $t \le \tau_p+\tilde{s}_p$, it is seen that $CK_p^4(\log(p)^{9/2}(p\widetilde{F}(t))^{1/2} = o(p\widetilde{F}(t)) = o(v_p(t))$. Inserting this into (5.85) gives that $|(\tilde{V}_p(t)-v_p(t))/v_p(t)| = o(1)$ with probability at least $1-o(1/p)$. By similar argument, $|\sqrt{n_p}(\tilde{M}_p(t)-m_p(t))/v_p(t)| = o(1)$ with probability at least $1-o(1/p)$. Combining these gives (5.84).

We now proceed to show the lemma in the second case. Let $h(t) = \hat{V}_p(t)/\tilde{V}_p(t)$. Write

$$\sqrt{n_p}|Sep(t,Z,\mu,\hat{\Omega})-Sep(t,\tilde{Z},\mu,\Omega)| \le \sqrt{1/h(t)}\cdot I+|\sqrt{1/h(t)}-1|\cdot II, \tag{5.86}$$

where $I = \sqrt{n_p}|\hat{M}_p(t)-\tilde{M}_p(t)|(\tilde{V}_p(t))^{-1/2}$ and $II = \sqrt{n_p}\tilde{M}_p(t)(\tilde{V}_p(t))^{-1/2}$. Recall that by Lemmas 5.4 and 5.12, $\sqrt{n_p}m_p(t) \le K_p^2(\log p)^{3/2}p\widetilde{F}(t) \lesssim K_p^2(\log p)^{3/2}v_p(t)$. Using Lemma 5.4 and (5.83)-(5.84),

$$|h(t)-1| = \frac{p\widetilde{F}(t)}{v_p(t)}\frac{v_p(t)}{\tilde{V}_p(t)}|\tilde{V}_p(t)-\hat{V}_p(t)|(p\widetilde{F}(t))^{-1} \le L_p[p^{-\theta/2}+(\log p)^{5/2}(p\widetilde{F}(t))^{-1/2}],$$

$I \le L_p(p\widetilde{F}(t)/v_p(t))^{1/2}(p\widetilde{F}(t))^{-1/2}[p^{1-\theta/2}\widetilde{F}(t)+(p\widetilde{F}(t))^{1/2}] \le L_p\cdot[p^{-\theta/2}(p\widetilde{F}(t))^{1/2}+1]$, and

$$II \lesssim (p\widetilde{F}(t)/v_p(t))(p\widetilde{F}(t))^{-1}\sqrt{n_p}[|\tilde{M}_p(t)-m_p(t)|+m_p(t)] \le L_p.$$

Recall that $p\widetilde{F}(t) \ge K_p^{10}(\log p)^{10}$. This together with the inequality above for $h(t)$ ensures that $|h(t)-1| \le o(1)$. Inserting these into (5.86) gives

$$|Sep(t,Z,\mu,\hat{\Omega}) - Sep(t,\tilde{Z},\mu,\Omega)| \le L_p\cdot n_p^{-1/2}[p^{-\theta/2}(p\widetilde{F}(t))^{1/2}+1],$$

and the claim follows.

Similarly to Lemma 3.4, we see that if in addition $\Omega \in \widetilde{M}_p^*(a,b,K_p)$, then the term $L_p p^{\frac{1-\theta}{2}-\frac{1}{4}\max\{4\beta-2r,3\beta+r\}}$ in the upper bound of the claim can be removed using the same proof as above. This concludes the proof of the lemma. □

5.24. *Proof of Theorem 2.3.* Note that for any $0<x<1$,

$$\bar{\Phi}^{-1}(x) = \sqrt{-2\log x} + O\Big(\frac{\log\log(\frac{1}{\sqrt{2\pi\log x}})}{\log x^{-1}}\Big), \tag{5.87}$$

where the last term is negligible compared to the first term. So $\bar{\Phi}^{-1}(\text{misclassification}|t)$ can be well approximated by $E(t) \equiv \sqrt{-2\log\big(P(YL_t(X,\Omega)<0|t)\big)}$.

We write $\widetilde{Sep}(t) = \widetilde{Sep}(t,\epsilon_p,\tau_p,\Omega)$, $Sep(t) = Sep(t,\tilde{Z},\mu,\Omega)$, and $T_{ideal} = T_{ideal}(\epsilon_p,\tau_p,\Omega)$ for short. The following lemmas are proved in Section 5.25 and Section 5.26 respectively.

LEMMA 5.15. *Fix a constant $\kappa>0$. As $p\to\infty$, for any sequence $t_p \in (0,\tau_p+\tilde{s}_p]$ with $\tilde{s}_p$ defined in Lemma 5.4 such that $\widetilde{Sep}(t_p) \ge L_p p^{\kappa}$, we have $P(YL_t(X,\Omega)<0|t=t_p) = \bar{\Phi}\big((1+o(1))\frac{1}{2}\widetilde{Sep}(t_p)\big)$.*

LEMMA 5.16. *For any sequence of closed subset $A_p \subset [0,\tau_p+\tilde{s}_p]$ with $\tilde{s}_p$ defined in lemma 5.4, if there exists a constant $\kappa>0$ such that $\sup_{t\in A_p}\{\widetilde{Sep}(t)\} \ge p^{\kappa}$ for sufficiently large p, then*

$$\sup_{t\in A_p} E(t) \lesssim \frac{1}{2}\sup_{0<t\le\tau_p+\tilde{s}_p}\widetilde{Sep}(t).$$

Now we proceed to prove the theorem. The key is to show

$$\min_{t>\tau_p+\tilde{s}_p} P(YL_t(X,\Omega)<0|t) > \bar{\Phi}\Big(\big(1+o(1)\big)\frac{1}{2}\widetilde{Sep}\big(T_{ideal}\big)\Big), \tag{5.88}$$

$$\min_{0<t\le\tau_p+\tilde{s}_p} P(YL_t(X,\Omega)<0|t) = \bar{\Phi}\Big(\big(1+o(1)\big)\frac{1}{2}\widetilde{Sep}\big(T_{ideal}\big)\Big). \tag{5.89}$$

Then combining the above results completes the proof of the theorem.

We first prove (5.88). When $r < \beta$, by proof (1b) in Lemma 3.4 we have $Sep(t, \tilde{Z}, \mu, \Omega) \le L_p p^{\frac{1-\theta}{2} - \frac{1}{4}\max\{4\beta-2r, 3\beta+4\}}$ for all $t > \tau_p + \tilde{s}_p$ with probability at least $1 - o(p^{-1})$. When $r \ge \beta$, by Lemma 3.4 we have

$$Sep(t) = \widetilde{Sep}(t) + (Sep(t) - \widetilde{Sep}(t)) \le \widetilde{Sep}(t) + L_p p^{-\theta/2}.$$

Following the same lines as those in the proof of Lemma 2.2 we can show that for $r \ge \beta$, $\widetilde{Sep}(t) \le L_p p^{\frac{1-\theta}{2} - c_8}$ with $c_8 = c_8(\beta, r) > \delta(\beta, r)$ for all $t > \tau_p + \tilde{s}_p$. Combining these and recalling that $r > \rho^*_\theta(\beta)$ and $\beta \in (\frac{1-\theta}{2}, 1-\theta)$, we have $Sep(t) \le L_p p^{\frac{1-\theta}{2} - c_9(\beta,r)}$ with $c_9(\beta, r)$ some constant whose value depends on whether $r < \beta$ or $r \ge \beta$ and satisfies $c_9(\beta, r) > \delta(\beta, r)$, for all $t > \tau_p + \tilde{s}_p$, with probability at least $1 - o(p^{-1})$. Recall that $\widetilde{Sep}\big(T_{ideal}\big) = L_p p^{\frac{1-\theta}{2} - \delta(\beta,r)}$. Thus,

$$\begin{aligned} P(YL_t(X, \Omega) < 0|t) &= \bar{\Phi}\Big(\frac{1}{2} Sep(t)\Big) \ge \bar{\Phi}\Big(L_p p^{\frac{1-\theta}{2} - c_9(\beta,r)}\Big)(1 - o(p^{-1})) \\ &\gg \bar{\Phi}\Big(\big(1 + o(1)\big)\frac{1}{2}\widetilde{Sep}\big(T_{ideal}\big)\Big). \end{aligned}$$

This completes the proof of (5.88).

Next we prove (5.89). Since (5.87) ensures that $\bar{\Phi}^{-1}(\text{misclassification}|t)$ can be well approximated by $E(t)$, we only need to prove

$$\sup_{\{0 < t \le \tau_p + \tilde{s}_p\}} E(t) \le \frac{1}{2} \sup_{\{0 < t \le \tau_p + \tilde{s}_p\}} \widetilde{Sep}(t)(1 + o(1)). \tag{5.90}$$

Then, $\bar{\Phi}\big(\frac{1}{2}\sup_{0 < t \le \tau_p + \tilde{s}_p} \widetilde{Sep}(t)(1 + o(1))\big)$ provides a lower bound for the misclassification rate $P(YL_t(X, \Omega) < 0|t)$ for $0 < t \le \tau_p + \tilde{s}_p$. Taking $t_p = T_{ideal}$ in Lemma 5.15 and noting that $T_{ideal} \in (0, \tau_p + \tilde{s}_p]$ shows $P(YL_t(X, \Omega) < 0|T_{ideal}) = \bar{\Phi}\big(\frac{1}{2}\sup_{0 < t \le \tau_p + \tilde{s}_p} \widetilde{Sep}(t)(1 + o(1))\big)$. Combining these yields (5.89).

We now proceed to prove (5.90). Define $A_p = \{t : t \in (0, \tau_p + \tilde{s}_p],\ \widetilde{Sep}(t) \le \frac{1}{2}\sup_{0 < t \le \tau_p + \tilde{s}_p}\{\widetilde{Sep}(t)\}\}$. Then by Lemma 5.16 and (5.87), for large enough p,

$$\sup_{t \in A_p} E(t) \le (1 + o(1))\frac{1}{2}\sup_{0 < t \le \tau_p + \tilde{s}_p}\{\widetilde{Sep}(t)\}.$$

So it remains to show that uniformly for all $t \in A_p^c \equiv (0, \tau_p + \tilde{s}_p] \setminus A_p$,

$$E(t) \le (1 + o(1))\frac{1}{2} \sup_{0 < t \le \tau_p + \tilde{s}_p} \widetilde{Sep}(t). \tag{5.91}$$

We proceed to prove the above claim (5.91). Introduce the event

$$B_p = \{\sup_{t\in A_p^c} \frac{|\tilde{M}_p(t) - m_p(t)|}{m_p(t)} \le L_p p^{-\kappa}, \qquad \sup_{t\in A_p^c} \frac{|\tilde{V}_p(t) - v_p(t)|}{p\widetilde{F}(t)} \le L_p p^{-\kappa}\}.$$

where $\kappa = (1-\theta)/2 - \delta(\beta, r) > 0$. The proof has two steps: (a) show that $P(B_p) \ge 1 - o(1/p)$, and then (b) show that on the event B_p, the desired claim in the lemma holds.

We first show (a). Recall that we have proved in (2.13) that $\sup_{0<t<\sqrt{2\log p}} \widetilde{Sep}(t) = L_p p^{\kappa}$. By Lemma 5.4 and (5.78), $v_p(t) \ge Cp\widetilde{F}(t)$ with some constant $C > 0$, where the value of C depends on whether $r < \beta$ or $r \ge \beta$. Moreover, by definition of A_p^c, $\widetilde{Sep}(t) \ge \frac{1}{2}L_p p^{-\kappa}$ for $t \in A_p^c$. It follows that $m_p(t) = \frac{1}{2}\sqrt{v_p(t)}\widetilde{Sep}(t) \ge \sqrt{Cp\widetilde{F}(t)L_p p^{\kappa}}$. On the other hand, by Lemma 5.12 $m_p(t) \le L_p p^{1-\theta/2}\widetilde{F}(t)$, so we can derive $p\widetilde{F}(t) \ge L_p p^{2\kappa+\theta}$ and consequently, $\sqrt{n_p}m_p(t) \ge L_p p^{2\kappa+\theta}$ and $v_p(t) \ge L_p p^{2\kappa+\theta}$. By Lemma 5.14 and using similar arguments as those in Lemma 5.13, we can prove that for each $t \in A_p^c$,

$$P\Big(\frac{|\tilde{M}_p(t) - m_p(t)|}{m_p(t)} \ge L_p p^{-\kappa}\Big) \le o(\frac{1}{p^3}), \quad P\Big(\frac{|\tilde{V}_p(t) - v_p(t)|}{v_p(t)} \ge L_p p^{-\kappa}\Big) \le o(\frac{1}{p^3}).$$

Using the grid point method as that for proving (3.1) shows that $P(B_p) \ge 1 - o(1/p)$.

We now show (b). On the event B_p,

(5.92)
$$\tilde{M}_p(t)/\sqrt{\tilde{V}_p(t)} \le (1 + L_p p^{-\kappa}) m_p(t)/\sqrt{v_p(t)} \le (1 + L_p p^{-\kappa})\frac{1}{2} \sup_{0<t\le\tau_p+\tilde{s}_p} \widetilde{Sep}(t).$$

This together with the definition of $E(t)$ completes the proof of claim (b).

By (5.92), uniformly over all $0 < t \le \tau_p + \tilde{s}_p$,

$$\begin{aligned} P(YL_t(X,\Omega) < 0|t) &\ge \bar{\Phi}\Big((1 + L_p p^{-\kappa})\frac{1}{2} \sup_{0<t\le\tau_p+\tilde{s}_p} \widetilde{Sep}(t)\Big) P(B_p) \\ &\ge \bar{\Phi}\Big((1 + L_p p^{-\kappa})\frac{1}{2} \sup_{0<t\le\tau_p+\tilde{s}_p} \widetilde{Sep}(t)\Big)(1 - o(\frac{1}{p})). \end{aligned}$$

This, together with (5.87), proves (5.91) and completes the proof of the theorem. □

5.25. *Proof of Lemma 5.15.* Write for short $\widetilde{Sep}(t) = \widetilde{Sep}(t, \epsilon_p, \tau_p, \Omega)$, $\tilde{M}_p(t) = M_p(t, \tilde{Z}, \mu)$, $\tilde{V}_p(t) = V_p(t, \tilde{Z}, \Omega)$, $m_p(t) = m_p(t, \epsilon_p, \tau_p, \Omega)$, and $v_p(t) = v_p(t, \epsilon_p, \tau_p, \Omega)$. Define event

$$B_p = \{|\tilde{V}_p(t_p) - v_p(t_p)| \le L_p p^{-\theta/2} p\widetilde{F}(t_p),\ |\tilde{M}_p(t_p) - m_p(t_p)| \le L_p p^{-\theta/2} m_p(t_p)\}.$$

The key is to first show that (a)

$$P(B_p^c) \le \exp\Big(-\frac{1}{2}\log(p)(\widetilde{Sep}(t_p))^2 \cdot (1 + o(1))\Big), \tag{5.93}$$

and then show that (b) on the event B_p. Combining (a) and (b) proves the desired claim holds.

We first prove claim (a). Note that by Lemma 5.4, $v_p(t) \ge Cp\widetilde{F}(t)$ with some constant $C > 0$, where the value of C depends on whether $r \ge \beta$ or $r < \beta$. Further by Lemma 5.12, $0 < \sqrt{n_p} m_p(t) \le K_p^2 (\log p)^{3/2} p\widetilde{F}(t) \le CK_p^2 (\log p)^{3/2} v_p(t)$, and so that $\sqrt{n_p} m_p(t) \ge Cn_p m_p^2(t)/[K_p^2 (\log p)^{3/2} v_p(t)] = \frac{Cn_p}{K_p^2 (\log p)^{3/2}} (\widetilde{Sep}(t))^2$. Taking $\lambda_p = K_p (\log p) \Big(\frac{\sqrt{n_p}\widetilde{Sep}^2(t_p)}{c_2 m_p(t_p)}\Big)^{1/2} m_p(t_p)$, then $\lambda_p \le L_p m_p(t_p)$. It follows that $P\big(\sqrt{n_p}|\tilde{M}_p(t_p) - m_p(t_p)| \ge CK_p^3 (\log(p))^2 \cdot L_p m_p(t_p)\big) \le P(\sqrt{n_p}|\tilde{M}_p(t_p) - m_p(t_p)| \ge CK_p^3 (\log(p))^2 \lambda_p)$, where by Lemma 5.14, the right hand side

$$\le CK_p^3 (\log p)^2 \exp\Big(-\big(\widetilde{Sep}(t_p)\big)^2 (\log p)\Big).$$

Since $\widetilde{Sep}(t_p) \ge L_p p^{\kappa} \to \infty$, it follows easily that

$$P\big(|\tilde{M}_p(t_p) - m_p(t_p)| \ge L_p p^{-\theta/2} m_p(t_p)\big) \le \exp\Big(-\big(\widetilde{Sep}(t_p)\big)^2 (\log p)(1 + o(1))\Big). \tag{5.94}$$

Next we consider $\tilde{V}_p(t)$. Let $\lambda_p = \widetilde{Sep}(t_p)\sqrt{(\log p) K_p p\widetilde{F}(t_p)}$. Using the same technique as for proving (5.94) we obtain that $\lambda_p \le L_p p^{-\theta/2} p\widetilde{F}(t)$. Further, by Lemma 5.14 we have

$$P(|\tilde{V}_p(t_p) - v_p(t_p)| \ge L_p p^{-\theta/2} p\widetilde{F}(t_p)) \le \exp\Big(-\big(\widetilde{Sep}(t_p)\big)^2 (\log p)\big(1 + o(1)\big)\Big). \tag{5.95}$$

Combing (5.94) with (5.95) proves (5.93).

On the set B_p, since $v_p(t_p) \ge Cp\widetilde{F}(t_p)$ by Lemma 5.4, we have $\frac{\tilde{V}_p(t_p)}{v_p(t_p)} = 1 + o(1)$, $\frac{\tilde{M}_p}{m_p(t_p)} = 1 + o(1)$. Therefore,

$$\frac{\tilde{M}_p(t_p)}{\sqrt{\tilde{V}_p(t_p)}} = \frac{m_p(t_p)}{\sqrt{v_p(t_p)}}\Big(1 + o(1)\Big) = \widetilde{Sep}(t_p)\big(1 + o(1)\big). \tag{5.96}$$

Combining (5.93) with (5.96), the misclassification rate can be bounded as

$$P(YL_t(X,\Omega)<0|t_p)\le \bar{\Phi}\left(\frac{1}{2}\widetilde{Sep}(t_p)\big(1+o(1)\big)\right)+P(B_p^c)\lesssim \bar{\Phi}\left(\frac{1}{2}\widetilde{Sep}(t_p)\big(1-o(1)\big)\right),$$

and

$$P(YL_t(X,\Omega)<0|t_p)\ge \bar{\Phi}\left(\frac{1}{2}\widetilde{Sep}(t_p)\big(1+o(1)\big)\right)P(B_p)\gtrsim \bar{\Phi}\left(\frac{1}{2}\widetilde{Sep}(t_p)\big(1+o(1)\big)\right).$$

Thus the claim follows easily. □

5.26. *Proof of Lemma 5.16.* Recall that $P(YL_t(X,\Omega)<0|t)=\bar{\Phi}\big(\tilde{M}_p(t)/\sqrt{\tilde{V}_p(t)}\big)$. By (5.87), to prove the lemma, it suffices to prove that uniformly for all $t\in A_p$,

$$\tilde{M}_p(t)/\sqrt{\tilde{V}_p(t)}\le (1+o(1))\frac{1}{2}\sup_{t\in A_p}\widetilde{Sep}(t). \tag{5.97}$$

Write $\widetilde{Sep}(t)=\widetilde{Sep}(t,\epsilon_p,\tau_p,\Omega)$ for short. We consider the cases (a) $p\widetilde{F}(t)\ge K_p^7(\log p)^7$, $\sqrt{n_p}m_p(t)\ge K_p^7(\log p)^7$, (b) $\sqrt{n_p}m_p(t)\le K_p^7(\log p)^7$, $p\widetilde{F}(t)\ge K_p^7(\log p)^7$, and (c)$\sqrt{n_p}m_p(t)\ge K_p^7(\log p)^7$, $p\widetilde{F}(t)\le K_p^7(\log p)^7$ separately.

For case (a), define the event

$$B_p=\{\sup_{t\in A_p}\frac{|\tilde{M}_p(t)-m_p(t)|}{m_p(t)}\le\frac{1}{\sqrt{\log p}},\qquad \sup_{t\in A_p}\frac{|\tilde{V}_p(t)-v_p(t)|}{p\widetilde{F}(t)}\le\frac{1}{\sqrt{\log p}}\}.$$

We will first prove $P(B_p^c)\le o(1/p)$. Let $\lambda=\lambda_p=CK_p^{-3}(\log p)^{-5/2}\sqrt{n_p}m_p(t)$ with $C>0$ some constant. Then by Lemma 5.14, using similar arguments as those in Lemma 5.13 we obtain that with probability at least $1-o(p^{-3})$, $|\tilde{M}_p(t)-m_p(t)|\le(\log p)^{-1/2}m_p(t)$. Using the grid points method as that in Theorem 3.1, we can prove that except for a probability of $o(1/p)$,

$$\sup_{t\in A_p,\sqrt{n_p}m_p(t)\ge(\log p)^{19}}\frac{|\tilde{M}_p(t)-m_p(t)|}{m_p(t)}\le(\log p)^{-1/2}$$

As for $\tilde{V}_p(t)$, using similar argument and by Lemma 5.13 we obtain that with probability at least $1-o(1/p)$,

$$\sup_{t\in A_p,p\widetilde{F}(t)\ge(\log p)^{19}}\frac{|\tilde{V}_p(t)-v_p(t)|}{p\widetilde{F}(t)}\le(\log p)^{-1/2}. \tag{5.98}$$

Thus we have proved the desired claim that $P(B_p) \geq 1 - o(1/p)$.

Next by Lemma 5.4, $p\widetilde{F}(t)/v_p(t) \leq C$ for all $0 < t \leq \tau_p + \tilde{s}_p$. Then on the event B_p,

$$\frac{\tilde{M}_p(t)}{m_p(t)} = 1 + o\Big(\frac{1}{\sqrt{\log p}}\Big) \qquad \frac{\tilde{V}_p(t)}{v_p(t)} = 1 + o\Big(\frac{1}{\sqrt{\log p}}\Big)\frac{p\widetilde{F}(t)}{v_p(t)} = 1 + o\Big(\frac{1}{\sqrt{\log p}}\Big),$$

where the $o(1)$ is uniformly over all t. Therefore, for any $t \in A_p$,

$$\tilde{M}_p(t)/\sqrt{\tilde{V}_p(t)} = (1 + o(1))m_p(t)/\sqrt{v_p(t)} \leq (1 + o(1))\frac{1}{2}\sup_{t\in A_p}\widetilde{Sep}(t),$$

and (5.97) has been proved.

Now we consider case (b). By the proof of Lemma 5.13 we obtain that except for a probability of $o(1/p)$, for any $t \in A_p$, $\tilde{M}_p(t) \leq m_p(t) + L_p n_p^{-1/2} \leq L_p n_p^{-1/2}$. Since we assumed that $p\widetilde{F}(t) \geq K_p^7(\log p)^7$, by (5.98) and the same argument as that for (5.84), we have $\frac{\tilde{V}_p(t)}{v_p(t)} = 1 + o(\frac{1}{\sqrt{\log p}})$. Since by lemma 5.4, $v_p(t) \geq Cp\widetilde{F}(t) \geq C(\log p)^{-1/2}$ with some constant $C > 0$ whose value depends on whether $r \geq \beta$ or $r < \beta$. Thus, with probability at least $1 - o(1/p)$, for any $t \in A_p$,

$$\tilde{M}_p(t)/\sqrt{\tilde{V}_p(t)} \leq L_p n_p^{-1/2}/\sqrt{v_p(t)} \leq L_p n_p^{-1/2}. \tag{5.99}$$

Thus, the claim in the lemma follows automatically by the assumption that $\sup_{t\in A_p}\{\widetilde{Sep}(t)\} \geq p^\kappa$ with $\kappa > 0$.

Finally we consider case (c). By Lemma 3.1, $p\widetilde{F}_p(t) \leq L_p$. Thus, using the same arguments as those for proving Lemma 3.4, part (1b) we obtain that

$$\tilde{M}_p(t)/\sqrt{\tilde{V}_p(t)} \leq L_p n_p^{-1/2}. \tag{5.100}$$

Using similar arguments as in case (b), we prove that (5.97) continue to hold in case (c). This completes the proof of the lemma. □

6. Appendix.

6.1. *Proof of Lemma 5.12.* Recall that $\bar{\Phi} = 1 - \Phi$ is the survival function of $N(0, 1)$. The following lemma is proved below.

LEMMA 6.1. *For any $t > 0$ and $u > 0$, there are universal constants $C_1 > 0$ and $C_2 \geq 1$ such that $C_1 \min\{t, \frac{1}{u}\} \leq \frac{1}{u} \cdot \frac{\bar{\Phi}(t-u)-\bar{\Phi}(t+u)}{\Phi(t-u)+\Phi(t+u)} \leq C_2(1 + t)$.*

We now show Lemma 5.12. Let $\tilde{\mu} = \Omega\mu$ for short. First, by definitions, $\sqrt{n_p}m_p(t) = \sum_{j=1}^p E[\sqrt{n_p}\tilde{\mu}(j)\mathrm{sgn}(\tilde{z}(j))1\{|\tilde{Z}(j)| \geq t\}] = \sum_{j=1}^p E[\sqrt{n_p}\tilde{\mu}(j)(\bar{\Phi}(t - \sqrt{n_p}\tilde{\mu}(j)) - \bar{\Phi}(t + \sqrt{n_p}\tilde{\mu}(j)))]$. Noting that for any fixed $t > 0$, $u[\bar{\Phi}(t-u) - \bar{\Phi}(t+u)]$ is a symmetric function,

$$\sqrt{n_p}m_p(t) = \sum_{j=1}^p E[|\sqrt{n_p}\tilde{\mu}(j)|(\bar{\Phi}(t - |\sqrt{n_p}\tilde{\mu}(j)|) - \bar{\Phi}(t + |\sqrt{n_p}\tilde{\mu}(j)|))]. \tag{6.1}$$

Similarly, we have

$$n_p u_p(t) = \sum_{j=1}^p E[n_p\tilde{\mu}^2(j)(\bar{\Phi}(t - |\sqrt{n_p}\tilde{\mu}(j)|) + \bar{\Phi}(t + |\sqrt{n_p}\tilde{\mu}(j)|))]. \tag{6.2}$$

Since that Ω is K_p-sparse and that $|\sqrt{n_p}\mu(j)| \leq \tau_p \leq \sqrt{2\log(p)}$, $|\sqrt{n_p}\tilde{\mu}(j)| = \sum_{k=1}^p |\Omega(j,k)| \cdot |\sqrt{n_p}\mu(k)| \leq K_p\sqrt{2\log(p)}$. Comparing (6.1) and (6.2), the first claim follows by Lemma 6.1. The second claim follows easily from the first claim and that $|\sqrt{n_p}\tilde{\mu}(j)| \leq K_p\tau_p$. □

6.2. *Proof of Lemma 6.1.* Consider the first inequality first. Let $\phi(\cdot)$ be the density of $N(0,1)$. For any real number v, write

$$\frac{\bar{\Phi}(t-v)}{\phi(t-v)} = \frac{\int_0^\infty \phi(x + (t-v))dx}{\phi(t+v)} = \int_0^\infty e^{-(t-v)x}e^{-x^2/2}dx,$$

where the right hand side is strictly monotone in v. Therefore, $\bar{\Phi}(t-u)/\phi(t-u) \geq \bar{\Phi}(t+u)/\phi(t+u)$ or equivalently, $\bar{\Phi}(t+u)/\bar{\Phi}(t-u) \leq \phi(t+u)/\phi(t-u)$. Combining this with basic algebra,

$$\frac{1}{u}\left[\frac{\bar{\Phi}(t-u) - \bar{\Phi}(t+u)}{\bar{\Phi}(t-u) + \bar{\Phi}(t+u)}\right] \geq \frac{1}{u}\left[\frac{\phi(t-u) - \phi(t+u)}{\phi(t-u) + \phi(t+u)}\right] = \frac{t}{ut}\left[\frac{e^{tu} - e^{-tu}}{e^{tu} + e^{-tu}}\right]. \tag{6.3}$$

When $0 < ut \leq 1$, the right hand side $\geq t \cdot \inf_{0<x<1}\{\frac{1}{x}\frac{e^x - e^{-x}}{e^x + e^{-x}}\}$. When $ut \geq 1$, by the monotonicity of the function $(e^x - e^{-x})/(e^x + e^{-x})$, the right hand side $\geq (1/u)\cdot[(e^{tu} - e^{-tu})/(e^{tu} + e^{-tu})] \geq (1/u)\cdot[(e - e^{-1})/(e + e^{-1})]$. Letting $C_1 = \min\{\inf_{0<x<1}\{\frac{1}{x}\frac{e^x - e^{-x}}{e^x + e^{-x}}\}, (e - e^{-1})/(e + e^{-1})\}$ gives the claim.

Consider the second inequality. When $u > 1$, the claim follows trivially, so we consider the case $0 < u \leq 1$ only. By Taylor expansion, there is a constant $c_3 \geq 1$ such that

$$\frac{1}{u}\frac{\bar{\Phi}(t-u) - \bar{\Phi}(t+u)}{\bar{\Phi}(t-u) + \bar{\Phi}(t+u)} \leq \frac{2u\max_{\{t-u<s<t+u\}}\{\phi(s)\}}{\bar{\Phi}(t-u)} \leq c_3\frac{\phi(t-u)}{\bar{\Phi}(t-u)}, \tag{6.4}$$

where in the second inequality we have used $t > 0$ and $u < 1$. At the same time, By Mills' ratio [41], there is a constant $c_4 > 0$ such that $\bar{\Phi}(t) \le c_4 \cdot (t\phi(t))$. Therefore, $\phi(t-u)/\bar{\Phi}(t-u) \le c_4(1+|t-u|) \le 2c_4(1+t)$. Insert this into (6.4). The claim follows by letting $C_2 = \max\{1, 2c_3c_4\}$. □

6.3. *Proof of Lemma 5.3.* Note that $P(|X| \ge t, |Y| \ge t) = P(X \ge t, Y \ge t) + P(-X \ge t, Y \ge t) + P(X \ge t, -Y \ge t) + P(-X \ge t, -Y \ge t) \equiv I_1 + I_2 + I_3 + I_4$. Consider I_3. Define $\tilde{Y} = 2\tau - Y$. Then $(X, \tilde{Y})$ has joint normal distribution with mean $(0, \tau)$ and correlation $-\rho$. Since $\tau > 0$, it is seen that $I_3 = P(X \ge t, \tilde{Y} \ge t + 2\tau) \le P(X \ge t, \tilde{Y} \ge t)$. Similarly, we can obtain that $I_4 \le P(\tilde{X} \ge t, \tilde{Y} \ge t)$ with $\tilde{X} = -X$ and $\tilde{Y} = 2\tau - Y$. So we only need to bound I_1 and I_2.

Since the proofs are similar, we only show the case $\rho \ge 0$. Write $P(X \ge t|Y \ge t) = P(X \ge t, Y \ge t)/P(Y \ge t)$. First, by elementary calculus,

$$P(X \ge t, Y \ge t) \le \begin{cases} C\exp(-\frac{t^2}{2}), & (t-\tau) \le \rho t, \\ C\exp(-\frac{t^2 - 2\rho t(t-\tau) + (t-\tau)^2}{2(1-\rho^2)}), & (t-\tau) \ge \rho t. \end{cases}$$

Second, note that when $0 \le t \le \tau$, $P(Y \ge t) \ge 1/2$, and that when $t \ge \tau$, $P(Y \ge t) = \bar{\Phi}(t-\tau) \ge C[1+(t-\tau)]^{-1}\phi(t-\tau)$ (e.g. by Mills' ratio [41]), where we note that $[1 + (t-\tau)]^{-1} \ge (1+t)^{-1}$. Combining these with elementary algebra,

$$P(X \ge t|Y \ge t) \le \begin{cases} C\exp(-t^2/2), & 0 \le t \le \tau, \\ C(1+t)\exp(-\frac{t^2 - (t-\tau)^2}{2}), & \tau < t < \frac{\tau}{1-\rho}, \\ C(1+t)\exp(-\frac{((1-\rho)t + \rho\tau)^2}{2(1-\rho^2)}), & t > \frac{1}{1-\rho}\tau. \end{cases}$$

Since $0 \le \rho \le a$, the claim follows by basic algebra. □

6.4. *Proof of Lemma 5.9.* Write $h(t) = \bar{\Phi}(t)/\phi(t)$ for short. For positive functions $f(t)$ and $g(t)$ defined over $(0, \infty)$, we say that $f(t) \asymp g(t)$ if there are constants $C_2 > C_1 > 0$ such that $C_1 \le f(t)/g(t) \le C_2$ for all $t > 0$. The following claims can be proved by elementary calculus and Mills' ratio [41] so we omit the proof. (a) $h(t) \asymp C\min\{1, 1/t\}$, (b) $h'(t)/h(t) = t - 1/h(t)$ and $(t^{-1} - t^{-3}) < h(t) < (t^{-1} - t^{-3} + 6t^{-5})$, and (c) $h'(-t)/h(-t) \le -C\max\{1, t\}$ for all $t > 0$.

To show the lemma, it suffices to show that $m_2'(t) < 0$ for all $t > 0$. Write

$$m_2(t) = \frac{1}{h(t)} \frac{\bar{\Phi}(t-\tau_p) + \bar{\Phi}(t+\tau_p)}{\phi(t-\tau_p) + \phi(t+\tau_p)} \equiv \frac{1}{h(t)} \frac{h(t-\tau_p)\phi(t-\tau_p) + h(t+\tau_p)\phi(t+\tau_p)}{\phi(t-\tau_p) + \phi(t+\tau_p)}.$$

We show this for the case of $t \ge \tau_p$ and the case of $t < \tau_p$ separately.

Consider the first case. By direct calculations, it is seen

(6.5)
$$m_2(t) = \frac{1}{1+e^{-2\tau_p t}}[h(t-\tau_p)/h(t)] + \frac{e^{-2\tau_p t}}{1+e^{-2\tau_p t}}[h(t+\tau_p)/h(t)] \equiv m_{2a}(t) + m_{2b}(t).$$

Write for short $\xi(t) = h'(t-\tau_p)/h(t-\tau_p) - h'(t)/h(t)$. By (a)-(b) and direct calculations,

$$|m'_{2b}(t)| \le C\tau_p e^{-t\tau_p}, \qquad m'_{2a}(t) = \xi(t)[h(t-\tau_p)/h(t)] + O(\tau_p t e^{-\tau_p t}),$$

where we note $h(t-\tau_p)/h(t) \ge C$. Note that the claim follows trivially if $t \le \tau_p + 3$. Therefore, to show the claim, it is sufficient to show $\xi(t) \le -C\tau_p^{-1}\min\{1, (\tau_p/t)^2\}$ for all $t > \tau_p + 3$. Toward this end, note that by basic algebra and (b),

$$\xi(t) = -\tau_p - \frac{1}{h(t-\tau_p)} + \frac{1}{h(t)} \le -\tau_p - \frac{(t-\tau_p)}{(1-(t-\tau_p)^{-2}+6(t-\tau_p)^{-4})} + \frac{t}{1-t^{-2}}.$$

By basic algebra, we have that for sufficiently large τ_p and $t > \tau_p + 3$,

$$\xi(t) \le -(t-\tau_p)^{-1}\big[\frac{1-6(t-\tau_p)^{-2}}{1-(t-\tau_p)^{-2}+6(t-\tau_p)^{-4}}\big] + 1/t + 2t^{-3}.$$

The claim now follows from elementary calculus.

Consider the second case. Rewrite

$$m_2(t) = \frac{1}{[1+e^{-2\tau_p t}]h(t)}h(t-\tau_p) + \frac{e^{-2\tau_p t}}{1+e^{-2\tau_p t}}\frac{h(t+\tau_p)}{h(t)} \equiv m_{2c}(t)h(t-\tau_p) + m_{2d}(t),$$

and so

$$m'_2(t) = m'_{2c}(t)h(t-\tau_p) + m_{2c}(t)h'(t-\tau_p) + m'_{2d}(t).$$

Similarly, by (a)-(c),

$$|m'_{2d}(t)| \le C\tau_p^{-1},\ m'_{2c}(t) \le C,\ m_{2c}(t)h'(t-\tau_p) \le -C\max\{1,t\}\cdot\max\{1,(\tau_p - t)\}h(t-\tau_p).$$

Combining these gives

$$m'_2(t) \le C[-\max\{1,t\}\cdot\max\{1,(\tau_p - t)\} + C]h(t-\tau_p) + C.$$

Since $h(t-\tau_p) \ge C$, it is seen that $m'_2(t) < 0$ for sufficiently large τ_p and the claim follows.

The second claim $m_2(t) > 1$ follows directly from the first claim and $\lim_{t\to\infty} m_2(t) = 1$, which can be obtained immediately by (6.5). □

References.

[1] Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. Wiley, New York.

[2] Arias-Castro, E., Candes, E. and Plan, Y. (2011). Global testing under sparse alternatives: ANOVA, multiple comparisons, and the Higher Criticism. *Ann. Statist.* **39**, 2533-2556.

[3] Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations. *Bernoulli* **10**, 989-1010.

[4] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199-227.

[5] Breiman, L. (2001). Random forests. *Mach Learning* **24**, 5-32.

[6] Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **2**, 121-167.

[7] Cai, T., Jin, J. and Low, M. (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.* **35**, 2421-2449.

[8] Cai, T. and Liu, W. (2011). A direct estimation approach to sparse linear discriminant analysis. *J. Amer. Statist. Assoc.* **106**, 1566-1577.

[9] Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ^1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106**, 594-607.

[10] Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* 35, 2313-2351.

[11] Cayon, L., Jin, J. and Treaster, A. (2005). Higher criticism statistic: detecting and identifying non-Gaussianity in the WMAP first-year data. *Mon. Not. R. Astron. Soc.* **362**, 826-832.

[12] Dempster, A. P. (1972). Covariance selection. *Biometrics* **28**, 157-175.

[13] Dettling, M. and Buhlmann, P. (2003). Boosting for tumor classification with gene expression data. *Bioinformatics* **19**, 1061-1069.

[14] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32**, 962-994.

[15] Donoho, D. and Jin, J. (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA* **105**, 14790-14795.

[16] Donoho, D. and Jin, J. (2009). Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Phil. Trans. R. Soc. A* **367**, 4449-4470.

[17] Efron, B. (2009). Empirical Bayes estimates for large-Scale prediction problems. *J. Amer. Statist. Assoc.* **104**, 1015-1028.

[18] Fan, J., Feng, Y. and Tong, X. (2012). A road to classification in high dimensional space: the regularized optimal affine discriminant. *J. Roy. Statist. Soc. B* **74**, 745-771.

[19] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

[20] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179-188.

[21] Friedman, J., Hastie, T. and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432-441.

[22] Genovese, C., Jin, J., Wasserman, L. and Yao, Z. (2012). A comparison of the lasso and marginal regression. *J. Mach. Learn. Res.* **13**, 2107-2143.

[23] Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* **38**, 1686-1732.

[24] Hall, P., Pittelkow, Y. and Ghosh, M. (2008). Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. *J. Roy. Statist. Soc. Ser. B* **70**, 159-173.
[25] He, S. and Wu, Z. (2012). Gene-based Higher Criticism methods for large-scale exonic SNP data. *BMC Proceedings* **5**, S65.
[26] Ingster, Y. I. (1997). Some problems of hypothesis testing leading to infinitely divisible distribution. *Math. Methods Statist.* **6**, 47-69.
[27] Ingster, Y. I. (1999). Minimax detection of a signal for ℓ_n^p-balls. *Math. Methods Statist.* **7**, 401-428.
[28] Ingster, Y. I., Pouet, C. and Tsybakov, A. B. (2009). Classification of sparse high-dimensional vectors. *Phil. Trans. R. Soc. A* **367**, 4427-4448.
[29] Jager, L. and Wellner, J. A. (2007). Goodness-of-fit tests via phi-divergences. *Ann. Statist.* **35**, 2018-2053.
[30] Ji, P. and Jin, J. (2010). UPS delivers optimal phase diagram in high dimensional variable selection. *Ann. Statist.* **40**, 73-103.
[31] Jin, J. (2009). Impossibility of successful classification when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA* **106**, 8859-8864.
[32] Jin, J. and Wang, W. (2012). Optimal spectral clustering by higher criticism thresholding. *Working Manuscript.*
[33] Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24**, 1175-1182.
[34] Ravikumar, P., Wainwright, M. J., Raskutti, G. and Yu, B. (2011). High dimensional covariance estimation by minimizing ℓ^1-penalized log-determinant divergence. *Electron. J. Statist.* **5**, 935-980.
[35] Sabatti, C., Service, S. and Hartikainen, A. et al. (2008). Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* **41**, 35-46.
[36] Shao, J., Wang, Y., Deng, X. and Wang, S. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *Ann. Statist.* **39**, 1241-1265.
[37] Shorack, G. R. and Wellner, J. A. (1986). *Empirical processes with applications to statistics*. Wiley, New York.
[38] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. B* **58**, 267-288.
[39] Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* **99**, 6567-6572.
[40] Tukey, J. W. (1976). T13 N: The higher criticism. Course notes, Stat 411, Princeton University.
[41] Wasserman, L. (2006). *All of nonparametric statistics*. Springer, New York.
[42] Zhong, P., Chen, S. and Xu, M. (2012). Alternative tests to Higher Criticism for high dimensional means under sparsity and column-wise dependency. *Manuscript.*

Y. Fan
Information and Operations Management Department
Marshall School of Business
University of Southern California
Los Angeles, CA 90089
USA
E-mail: fanyingy@marshall.usc.edu

J. Jin
Department of Statistics
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
USA
E-mail: jiashun@stat.cmu.edu

Z. Yao
Section de Mathématiques
Ecole Polytechnique Fédérale de Lausanne
EPFL Station 8, 1015 Lausanne
Switzerland
E-mail: zhigang.yao@epfl.ch